

Perspektiven der Langzeitarchivierung multimedialer Objekte

Prof. Dr. Wolfgang Coy
Humboldt-Universität zu Berlin
Institut für Informatik



Perspektiven der Langzeitarchivierung multimedialer Objekte

Prof. Dr. Wolfgang Coy

Humboldt-Universität zu Berlin
Institut für Informatik

nestor – materialien 5

Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources

<http://www.langzeitarchivierung.de>

Projektpartner

Bayerische Staatsbibliothek, München

Bundesarchiv

Computer- und Medienservice / Universitätsbibliothek der Humboldt-Universität zu Berlin

Die Deutsche Bibliothek, Leipzig, Frankfurt am Main, Berlin (Projektleitung)

Generaldirektion der Staatlichen Archive Bayerns, München

Institut für Museumskunde, Berlin

Niedersächsische Staats- und Universitätsbibliothek, Göttingen

© 2006

nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit
Digitaler Ressourcen für Deutschland

Der Inhalt dieser Veröffentlichung darf vervielfältigt und verbreitet werden, sofern der Name des Rechteinhabers „nestor - Kompetenznetzwerk Langzeitarchivierung“ genannt wird. Eine kommerzielle Nutzung ist nur mit Zustimmung des Rechteinhabers zulässig.

Betreuer dieser Veröffentlichung:

Computer- und Medienservice der Humboldt-Universität zu Berlin
(Dr. Peter Schirnbacher)

URN: urn:nbn:de:0008-20051214015
<http://nbn-resolving.de/urn:nbn:de:0008-20051214015>

Aus der Sicht einer Einrichtung, die ganz praktisch die Probleme beim Umgang mit multimedialen Daten zu bewältigen hat, spiegelt die im Resümee der Expertise ausgesprochene Einschätzung zur Perspektive der Langzeitverfügbarkeit medialer Daten exakt die Realität wieder. Es gibt aus heutiger Sicht keine gesicherten Technologien, die bereits jetzt einsetzfähig wären, um die volle Originalität und Funktionalität eines digitalen Objekts auch für eine zukünftige Benutzung zu garantieren.

Daher ist es umso wichtiger, dass Institutionen, die sich dieser Aufgabe stellen wollen oder müssen, Empfehlungen an die Hand bekommen, die Ihnen sicher nicht in jedem Detail erklären, was zu tun ist. Die aber die Schwierigkeiten bei der digitalen Langzeitarchivierung multimedialer Daten darstellen und Strategien aufzeigen, die aus heutiger Sicht geeignet sind, eine zukünftige Nutzung der digitalen Daten zuzulassen.

Die Arbeit konzentriert sich auf den Bereich „Bibliothek“ und den Vergleich mit konventionellen Herangehensweisen. Die Darstellung der technischen Problemlage und der strategischen Überlegungen lässt sich ebenso auf andere Gebiete übertragen, deren Aufgaben um die Aufbewahrung digitaler Objekte wachsen.

Sicher wirkt die vorliegende Ausarbeitung an einigen Stellen äußerst desillusionierend, besonders, wenn sie Aussagen auf der Basis eines unendlichen Zeithorizonts trifft. Sie stellt damit auch die heutige Praxis bei der Produktion digitaler Medien in Frage. Die Archivierungsqualität wird bereits beim Produktionsprozess sei es durch die Entscheidung für ein bestimmtes Speicherformat oder für die Nutzung zusätzlicher Verfahren, wie z.B. digitale Signaturen oder Digital Rights Management Systemen entscheidend beeinflusst. Der Autor sieht in der Nutzung offener Standards und offener Formate einen Weg, um die digitale Langzeitarchivierung, insbesondere den Einsatz von Archivierungsstrategien wie Migration oder Emulation, auch wirtschaftlich handhabbar zu gestalten.

Die Arbeit besitzt den Charakter eines allgemeinen Überblicks. Damit erfüllt sie die Funktion, das Problembewusstsein zu schärfen. Um alle „Best Practise“- Lösungen zu erfassen und die Erfahrungen daraus zur Verfügung zu stellen und praktisch einsetzbar zu machen, hat nestor inzwischen die Arbeitsgruppe „media“ gebildet.

Die Komplexität der Aufgabe multimediale Daten für die Ewigkeit benutzbar zu halten ist zu groß, als dass sie von Einzelnen allein gelöst werden kann. Jeder Baustein, der uns der Lösung näher bringt, ist ein Beitrag zur Bewältigung dieser hochgradig kooperativen Aufgabe. In diesem Sinne verstehen wir die vorliegende Expertise als Beitrag zum großen Puzzle „nestor - Kompetenznetzwerk Langzeitarchivierung“.

Für die Partner des Projektes nestor – Kompetenznetzwerk Langzeitarchivierung

Dr. Peter Schirnbacher und Susanne Dobratz

Humboldt- Universität zu Berlin

Aufgabenstellung der Expertise

Der Begriff „multimedial“ umfasst im Sinne der Aufgabenstellung alle nicht-textbasierten medialen Objekte vom Bild über Audio, Video zu spezifischen Techniken der Visualisierung und Animation und komplexen, zusammengesetzten Anwendungen.

Bei diesen Objekten steht die Archivierung vor dem Problem eines besonders hohen Innovationszyklus, einer Vielzahl technischer Formate und unzureichender internationaler Normierung. Welche Entwicklungen sind zu erwarten, erhöht sich die Komplexität oder werden sich Standards durchsetzen? Welche Strategien sollen eingeschlagen werden, um multimediale Datenbestände mit vertretbarem Aufwand langfristig zugänglich halten zu können?

Ziel dieser Expertise ist es, Betreibern von multimedialen Archiven Empfehlungen (Eckpunkte, Prognosen, Rahmenbedingungen) an die Hand zu geben, die ihnen die Entwicklung von Strategien zur effektiven und nachhaltigen Archivierung unter den jeweiligen medialen Aspekten ihrer Sammlung ermöglichen. Dabei ist zu berücksichtigen, dass unterschiedliche Prioritäten in der Bewertung der Teilaspekte der Expertise zu verschiedenen Strategien führen dürften.

Bei der Bearbeitung der Problematik sollen folgende spezifische Fragestellungen betrachtet werden:

1. Medientypen: Erstellung eines Überblicks über die potenziell zu archivierenden medialen Techniken, die in Forschung und Lehre eingesetzt werden oder zukünftig eingesetzt werden könnten. Sind grundsätzlich neue Anforderungen für die Speicherung bzw. Archivierung zu erkennen?
2. Technische Formate: Erstellung eines Überblicks über die für Forschung und Lehre relevanten technischen Formate, deren Herkunft und Geschichte, Verbreitung sowie spezifische Eigenschaften (z.B. Ressourcenbedarf). Wie verhält sich die Steigerung der technischen Leistungsfähigkeit zur Kompatibilität (intern und extern)?
3. Nachhaltigkeit: Welche Formate lassen eine hohe Persistenz erwarten, welche eine gute Konversionsfähigkeit für künftige Migration? Welche Entwicklungen sind zu erwarten?
4. Definitionsmacht: Welche Durchsetzungsperspektiven haben offene Standards gegenüber firmendominierten, ggf. nicht-öffentlichen Entwicklungen?
5. Metadaten: Welche Bedeutung haben Metadaten für die jeweilige Konversionsfähigkeit? Auf welchem Stand ist die entsprechende Metadatenstandardisierung (Aufstellung relevanter Standards und evtl. konkurrierender Modelle zur Beschreibung multimedialer, auch komplexer Objekte)?
6. Organisation: Lassen sich zu erwartende Aufgaben sinnvoll zwischen Institutionen teilen? Ist Know-how zu einzelnen Bereichen etwa eher zentral oder verteilt aufzubauen? Ist es sinnvoll, Datenbestände unter technischen Gesichtspunkten zusammenzufassen?
7. Zeithorizont: Inwiefern beeinflussen unterschiedliche Bezugsrahmen von „Langzeit“ (z.B. 5, 10, 15 Jahre bzw. grundsätzlich offen) die vorliegende Expertise bzw. geeignete Archivierungsstrategien?
8. Strategiediskussion: Welche Faktoren müssen bei der Entwicklung einer Strategie berücksichtigt werden? Wie sind sie zu gewichten?

WOLFGANG COY

Perspektiven der Langzeitarchivierung multimedialer Objekte

Berlin 10.4.2005

Anschrift des Autors

Prof. Dr. Wolfgang Coy
Humboldt-Universität zu Berlin
Institut für Informatik
Unter den Linden 6
10099 Berlin

coy@hu-berlin.de

Inhaltsübersicht

Inhaltsübersicht	3
Kurzfassung der Empfehlungen.....	5
I. Die Digitale Bibliothek	9
II. Der digitale Bestand – Medientypen.....	15
III. Speichermedien und Speichermaterialien für multimediale Artefakte	23
IV. Technische Speicherformate – vom Text zum Multimediaobjekt	38
V. Emulation zur langfristigen Bereitstellung von Multimediadokumenten und Programmen	58
VI. Migrationspfade zur Bestandssicherung: Organisation, Zeithorizont, strategische Aspekte	62
Zusammenfassung und Empfehlungen.....	67
Referenzen	73

Kurzfassung der Empfehlungen

- Während die Speicherung und Nutzung multimedialer digitaler Objekte über vier bis acht Jahre relativ unproblematisch ist und für entsprechend aufbereitete Text- und Bilddaten Techniken zur zehn- und zwanzigjährigen Speicherung und Nutzung verfügbar sind, zeichnen sich zur Speicherung und Nutzung multimedialer Daten über hundert Jahre sich keine gesicherten Strategien ab.
- Bei selbsterzeugten oder nach eigenen Regeln, „proaktiv“ erzeugten Daten bestehen Chancen auf eine sachgerechte, ökonomisch und organisatorisch vertretbare langfristige Speicherung und Nutzung. Bei fremdbestimmten Daten sind keine solchen Wege gesichert. Die proaktive Sammlung digitaler Objekte ist deshalb, soweit möglich, zu fördern.
- Es werden auch künftig keine dauerhaften digitalen Speichermedien bzw. Materialien zur Verfügung stehen. Die Haltbarkeiten sind gegenüber Papierbeständen erheblich verkürzt. Notwendigerweise muß Bestandserhaltung bei digitalen Objekten durch dynamische Speicherung erfolgen: Programme und Daten müssen immer wieder auf neue Medien umkopiert werden und Programme müssen gegebenenfalls umkodiert werden, um sie an neue Geräte und Betriebssysteme anzupassen.
- Umkopieren und vor allem Umkodieren zur Anpassung an neue Geräte und Betriebssysteme wird zu einem beträchtlichen Kostenfaktor für digitale Bibliotheken.
- Neben der Objektspeicherung ist die Erstellung und Pflege zugehöriger Metadaten von wesentlicher Bedeutung. Metadaten sind frühzeitig zu erstellen und zeitnah zu pflegen.
- Authentizität und Integrität von digitalen Objekten müssen technisch und organisatorisch gesichert sein.
- Unter dem Aspekt der technologischen Erhaltung sind rechtlich fixierte und vollständig offen gelegte Standards gegenüber proprietären Systemen (auch so genannten Industriestandards) vorzuziehen. Bei Computerprogrammen ist der Zugriff auf den Quellcode höchst wünschenswert. Open-Source-Produktionen sind vorzuziehen, da sie das Umkodieren zur Nutzung unter neuen Geräten und Betriebssystemen erleichtern.
- Digitale Objekte sind technisch einfach kopierbar. Organisatorisch ist zur Datensicherung eine geeignete Backup-Strategie zu entwickeln, die jedes

Objekt mindestens in drei Kopien, auf unterschiedlichen Speichermedien, speichert.

- Abhängig von der Bedeutung des zu bewahrenden Objekts ist die Kopie auf technisch verschiedenen Datenträgern, gegebenenfalls auch in analoger Form, zu erwägen.
- Emulatorprogramme für Geräte und Betriebssysteme deuten einen Erfolg versprechenden Weg zur Bewahrung und Nutzung digitaler Objekte an.
- Bei digitalen Objekten lässt sich die reine Datenspeicherung und Datensicherung von den originären Aufgaben der Sammlung, der Bestandserschließung und der inhaltlichen Bestandssicherung trennen. Insbesondere unter dem Aspekt breitbandiger Rechnernetze könnte die Datensicherung auch an eine vertrauenswürdige außen stehende Stelle, etwa an ein Rechenzentrum, übergeben werden oder im Bibliotheksverbund betrieben werden.
- Besondere Probleme entstehen beim ausschließlichen Netzzugriff. Unter Langfristüberlegungen ist einer Speicherung beim Verlag unbedingt eine Speicherung im Haus oder die Hinterlegung bei einer vertrauenswürdigen dritten Stelle vorzuziehen.
- Rechtliche Regelungen können technische Möglichkeiten aushebeln. Dies gilt insbesondere bei Regelungen zum Schutz geistigen Eigentums, wie dem Umgehungsverbot eines „wirksamen Kopierschutzes,“ der angesichts der mangelhaften Qualität digitaler Speichermedien eine vorhersehbare Vernichtung gespeicherter Daten zulässt.

I want to build
a clock that ticks
once a year.

The century hand
Advances once every
100 years, and the
cuckoo comes out
on the millennium.

I want the cuckoo
to come out every
millennium for the
next 10,000 years.

If I hurry, I should
finish the clock
in time to see the
cuckoo come out
for the first time.

Danny Hillis, The Millennium Clock, *Wired*, 1995.

I. Die Digitale Bibliothek

„Erwerben, Sammeln Speichern, Bereitstellen“: Grundaufgaben der Bibliothek gelten auch im Übergang vom bisherigen Sammelgut zu digitalen Formen. Kann man daraus schließen, daß sich eigentlich nichts Wesentliches geändert hat? Worin unterscheidet sich die „Digitale Bibliothek“¹ von den herkömmlichen Formen? Gibt es überhaupt eine „Digitale Bibliothek,“ die sich von den tradierten Formen absetzt?

Digitale Bibliotheken in Reinform werden nur selten neu entstehen. Die typische Entwicklung ist der Erwerb digitalen Sammelguts mit der langfristigen Aufgabe zur Bestandwahrung und zur Bereitstellung. Auf den ersten Blick scheint digitales Sammelgut wie z.B. eine CD pflegeleicht: Ein stabiles HiTech-Objekt, das schon durch die Bauform vor Schmutzeinwirkungen geschützt wird, das Nässe oder Kratzer verträgt, von einer eigenen Hülle geschützt wird und für das einfach ein Backup erzeugt werden kann. Noch vorteilhaft scheint es, wenn das Sammelgut im Intranet oder Internet bereitgestellt, den Nutzern also nur als Datenstrom angeboten wird. Es entsteht eine optimale Bibliothekssituation: *Die Benutzer sind bei der Nutzung vom Sammelgut elektronisch getrennt.*

Leider wird mit dieser Trennung von Sammelgut und Nutzung das Grundproblem der langzeitigen Speicherung nicht wirklich gelöst. Jeff Rothenberg hat das Dilemma der Langfristspeicherung mit dem einfachen Satz beschrieben: *„Digital Information lasts forever...“*, aber er fügt auch hinterlistig hinzu *„or five years – whatever comes first.“* Die Ausgangslage ist widersprüchlich: Einerseits ist es noch nie so leicht gewesen, ein

¹ Das Wort „Digitale Bibliothek“ wird auch als Name einer CD-ROM-Edition von der *Berliner Directmedia Publishing GmbH* verwendet. Hier ist selbverständlich immer der generische Begriff gemeint.

bibliothekarisches Sammelobjekt zu lagern, nahezu perfekt zu kopieren und elektronisch zu verteilen. Texte, Grafik, Fotos, Sprache, Musik, Video oder Multimedialprogramme können digital in einheitlicher Weise als Binärströme gespeichert und übertragen werden und sie sind mit digitalen Prozessoren, also Computern, ohne großen Aufwand über Netze oder mittels Wechselspeichermedien zu verteilen und zu kopieren. Andererseits sind viele Fälle bekannt, wo digitale Speichermedien nicht mehr oder nur noch mit größten Schwierigkeiten gelesen, kopiert oder verarbeitet werden können. Diese Schwierigkeiten mögen unterschiedlichste Gründe haben: Speichermedien haben eine beschränkte Haltbarkeit, Lesegeräte veralten technologisch, Speicherformate sind nicht hinreichend dokumentiert oder Programme sind mittels aktuell verfügbarer Geräte und Betriebssysteme nicht mehr ausführbar, also technologisch veraltet. Während derartige Datenverluste bei Büro-PCs eine alltägliche Erfahrung sind, erstaunt es doch immer wieder, daß Datenverluste selbst bei Organisationen wie der NASA genannt werden, wo mit Millionenaufwand erstellte Satellitendaten nach wiederholten Presseberichten nicht mehr rekonstruiert werden können. Und selbst in der Bilddatenbank CORBIS, deren einziger Zweck die Bereitstellung gesammelter digitalisierter Bilddaten ist, wurden nach einem Bericht der *New York Times* vom 7.4.1998 einige Bilder auf Grund fehlerhafter Speichermedien verloren².

Selbst wenn digitale Daten durch eine Vielzahl von Kopien gesichert scheinen, bleibt ein deutliches Unbehagen. Es ist klar, daß die Göttinger Gutenberg-Bibel noch lesbar (aber vielleicht nicht mehr zugänglich) sein

² Stephen Manes, Time and Technology Threaten Digital Archives..., *New York Times*, 7.4.1998

wird, wenn ihre CD-Version aus technischen Gründen nicht mehr geöffnet werden kann. Andererseits ist sie durch ihre CD-ROM-Fassung erst einer breiten Öffentlichkeit umfassend zugänglich gemacht worden. Der Vatikan hat erst vor wenigen Jahren eine CD seiner Gemälde produziert, die nur unter Windows 3.11 aber nicht unter neueren Versionen lesbar ist. Für die Käufer der CD heißt dies, dass sie eine neue Version nachkaufen müssen – oder ihren Kauf als praktisch verloren ansehen müssen.

Die Frage ist also, welcher Aufwand zur Bestandssicherung digitaler Objekte zu treiben ist – und ob dies überhaupt Aussicht auf Erfolg hat. Und es stellt sich die Frage, wie weit sich diese Herausforderungen der Digitalen Bibliothek von bisherigen Aufgaben der Sammlung, Bestandssicherung oder Bereitstellung unterscheiden. Was ist also *neu* an der Digitalen Bibliothek?

- Digitale Objekte können die digitalen Äquivalente von Medien aller Art speichern. Sie stellen Texte in formatierter oder unformatierter Form bereit, sie bilden formatierte Datenbestände in Datenbanken oder Tabellenkalkulationsprogramme ab, ihre Daten können als Grafiken oder Fotos, Töne, Musik, Sprache oder Videos interpretiert, dargestellt, kopiert oder verteilt werden. Die Grundstruktur ist stets die gleiche: Ein Bitstrom, eine Folge von Signalen, die als Nullen oder Einsen interpretiert werden. Dieser Bitstrom kann einheitlich gespeichert, in Rechnern verarbeitet, mittels Speichermedien oder über Netze verteilt werden. Zur auffälligsten Eigenschaft digitaler Bitströme gehört: Sie sind unter Einsatz geeigneter Fehlerkorrekturverfahren nahezu perfekt kopierbar.

- Die Gleichförmigkeit der Bitströme erlaubt es, eine einheitliche digitale Speicherstruktur unabhängig von der medialer Form und dem Inhalt des gespeicherten Materials zu nutzen
- Mit der Digitalisierbarkeit unterschiedlichster medialer Formen weitet sich das potentielle bibliothekarische Sammelgut aus auf alle digitalen Objekte einschließlich Computerprogrammen, multimedialen Produktionen oder Computerspielen.
- Digitale Medien werden derzeit in Form von DVDs, Audio-CDs und CD-ROMs in einer Mischform vertrieben: Verpackungen, die neben einem Speichermedium Papierausdrucke oder *Booklets* enthalten. Diese verpackten Medien lassen sich ähnlich wie Bücher, Videobänder oder anderes herkömmliches Sammelgut behandeln oder aber über einen Netzzugriff bereitstellen.
- Doch materielle Trägermedien wie CD und DVD sind sehr wahrscheinlich nur noch eine Übergangsform, deren Einsatzzeit höchstens durch rechtliche Restriktionen verlängert wird. Auch dies wird durch geeignete Methoden des Digital Rights Managements (DRM) beim Netzzugang ändern.
- Die Perspektive der Digitalen Bibliothek liegt offensichtlich im Netz – als Intranet oder als Internetangebot. Damit streifen die digitalen Medien ihre Verbindung zum gedruckten Material einschließlich der Hülle ab. Mit der Netzspeicherung wird die bibliothekarische Sammlung (endlich?) einheitlich speicherbar und die bibliothekarische Pflege

dieser Speicherdaten kann vollständig von ihrer Lagerung getrennt werden. *Sie kann damit sogar aus der Bibliothek ausgelagert werden, ohne daß deren Kernkompetenz wesentlich berührt wäre.*

- Eine solche Trennung kann im Auftrag der Bibliothek geschehen, etwa durch die Auslagerung elektronisch gespeicherter Dissertationen und anderer wissenschaftlicher Publikationen in die Speicherarchitektur eines Rechenzentrums – oder durch die komplette Auslagerung zu einem externen Dienstleister. Bei universitären Bibliotheken ist Kooperation mit dem universitären Rechenzentrum nahe liegend; sie kann sogar als ein Ausgangspunkt zum Aufbau eines universitären I&K-Zentrums betrachtet werden.
- Ausgelagerte Speicherung kann aber auch gegen den Wunsch der Bibliothek durch eine Verlagsstrategie erzwungen werden, wenn etwa der Zugang zu wissenschaftlichen Zeitschriften nur noch vom Verlags-server aus erfolgen darf. Die Bibliothek wird derart zum Informationsmakler, der eine unmittelbare Informationsbereitstellung vornehmen kann – auf Widerruf. Originäre Aufgaben der Sammlung, Bestandserhaltung und dauerhaften Bereitstellung werden u.U. der Bibliothek abhängig von den vertraglichen Abreden entzogen. Dies kann zu einer äußerst problematischen Entwicklung führen.
- Bibliotheken können mit der Digitalisierung von medialer Produktion und Distribution neue Aufgaben zuwachsen. Die Sammlung digital-elektronischer Medien kann direkt übernommen werden, etwa im Rahmen eines Repositories der wissenschaftlichen Literatur, die in

ihrem Einzugsbereich erstellt wurde, im Rahmen eines „*self archiving*“-Konzeptes oder durch die Herausgabe elektronischer Zeitschriften oder Buchreihen – mit der realen Perspektive der Herstellung gedruckten Materials im Rahmen eines *Publishing on Demand*-Verfahrens. Auch hier sind die örtlichen Bedingungen entscheidend.

Festzuhalten ist: Die künftige Digitale Bibliothek wird als Ergebnis technischer Möglichkeiten und Umgestaltungen entstehen. Ihre Perspektive liegt in der Bereitstellung digitaler Materialien im Netz. Materielle Trägermedien wie CD und DVD sind sehr wahrscheinlich nur noch eine Übergangsform. Damit verbunden sind umfassende organisatorische, finanzielle und rechtliche Herausforderungen. Aufgaben der Nutzerberatung und Nutzerbetreuung müssen dieser Entwicklung angepaßt werden. Aufgaben der Sammlung und Bereitstellung einerseits und der Speicherung andererseits müssen in der digitalen Bibliothek nicht zwingend zusammenfallen.

II. Der digitale Bestand – Medientypen

Während es die klassische Bestandssicherung über weite Bereiche mit vorgegebenen Inhalten und Medien zu tun hat, verhalten sich digitale Objekte plastischer. Solange dem keine (lizenz-) rechtlichen oder technische Umstände entgegenstehen und der Aufwand vertretbar bleibt, lassen sich digitale Objekte umformen und in Grenzen den Anforderungen des digitalen Bestandes anpassen. Als mögliches Ziel könnte eine vollständige Migration der digitalen Materialien in den nicht mehr ortsgebundenen Netzzugriff gesehen werden – abhängig von den Aufgaben und Lizenzen als Intranet oder als Internetangebot.

Damit wird der Unterschied zwischen selbst erstelltem oder nach Vorgaben erstelltem und von der Bibliothek unbeeinflusst fremd erstelltem Material entscheidend. Bei Ersterem können Fragen der Sammlung, des Zugriffs und der Bestandserhaltung frühzeitig und zielorientiert behandelt werden, bei Letzterem ist dies nur in den Grenzen vorhandener Angebote möglich, bei denen typischerweise bloß die zeitnahe Bereitstellung als Aufgabe gesehen wird, aber Fragen bibliothekarischer Erschließung und Bestandserhaltung kaum interessieren. Zum langfristigen Speichern sind aber frühzeitige und dauerhafte Eingriffsmöglichkeiten bei der Erstellung und eventuellen technischen Wandlungen wie Umkopieren und Umkodieren von großer Bedeutung. Das gleiche gilt von der technischen und organisatorischen Kontrolle über Speicherung, Präsentation und Distribution. Bibliotheken geraten so in einen prinzipiellen Nachteil bei der Sammlung fremd bestimmte erstellter Materialien, der bislang in der Buchwelt kaum zum Tragen gekommen ist.³

³ ...wenn man von der Verwendung schlechter Papiere und Klebstoffe absieht.

Selbsterstellte Dokumente einschließlich solcher, bei denen die Autoren oder Verwertern expliziten Vorgaben der Bibliothek folgen, können dagegen mit einer klaren Vorstellung künftiger Nutzung generiert und gestaltet werden. Dissertationen und wissenschaftliche Arbeiten, Ausstellungskataloge, Materialsammlungen, aber auch alle anderen Veröffentlichungen in einem bibliothekseigenen oder bibliotheksnahen Verlag gehören in diese Kategorie. Bibliotheken können auch gemeinsame Sammlungen (*Repositories*) für die eigenständige Bereitstellung digitaler wissenschaftlicher Materialien im Rahmen des „Self Archiving“ nach eigenen Vorgaben definieren. Derart selbsterstellte Dokumente sind in Bezug auf die künftige Verwendung nach entsprechenden Standards auf geeigneten Speichermedien mit präzisen Metadaten herstellbar und einsetzbar. Bedingung ist freilich eine entsprechende Schulung und Kontrolle des Erstellungsprozesses. Sofern dies befriedigend gelingt, können Bestandserschließung und Bestandssicherung eine neue, bislang ungekannte Qualität erreichen. Dies kann zum herausragenden Kennzeichen digitaler Bibliotheken werden kann.

Bei fremderstellten Dokumenten wird somit die Bedeutung übergreifender Vereinbarungen deutlich. Dies verlangt typischerweise die Einigung auf Austauschformate, über die verwendete kompatible Software und Hardware, über die Speicherung im Netz, über die Lieferung von externen Metadaten, sowie die technische und rechtliche Möglichkeit des Umkopierens und Umkodierens. Bei vielen kommerziellen Produktionen ist dies nicht oder nur rudimentär gegeben. Ein Bewußtsein für die Notwendigkeit solcher Vereinbarungen ist bei den wenigsten Produzenten und Verwertern digitaler Objekte vorhanden, so dass die Protagonisten der

Digitalen Bibliothek hier noch viel Aufklärung unter den Aspekten der Bestandserschließung und Bestandserhaltung betreiben müssen.

Besondere Aufmerksamkeit ist dabei „offenen“ oder „freien Standards“ zu widmen, die sowohl die Offenlegung der Spezifikation oder der Quellen als auch eine frei verfügbare Lizenzierung betreffen können. „Offenheit“ oder „Freiheit“ zeigt sich dabei auf unterschiedlichen Ebenen. Offene Standards der Formate erleichtern die einheitliche Behandlung der gesammelten Objekte, vor allem aber eine einheitliche Migration bei der Bestandssicherung. Offene Inhalte betreffen die Lizenzierung des Urheberrechts, wie in der *Creative Commons Lizenz* oder bei Programmen in den Varianten der *Open Source Lizenzen* (wie beispielsweise der GNU Public License oder der BSS License). Im wissenschaftlichen Bereich sind offene Sammlungen und Archive (*Open Archives*) von wachsender Bedeutung. Eine gemeinsame, möglichst standardisierte Metadatenstruktur erlaubt eine vereinfachte und deutlich erweiterte digitale Bestandserschließung – wäre ein sichtbarer Vorteil Digitaler Bibliotheken.

Die Digitale Bibliothek hat einen gegenüber der herkömmlichen Bibliothek erweiterten Bestandsbegriff, der immer wieder mit dem Sammelauftrag abzugleichen ist. Der Bestand umfasst alle digitalen Varianten von Büchern, Zeitschriften, Grafik, Foto, Karten, Musikalien und anderen Druckmaterialien über Audio als Sprache, Musik und anderen Tondokumenten sowie Video und anderen Formen des Bewegtbildes bis zu Computerprogrammen, Computerspielen und multimedialen Produktionen aller Art. Wie weit dies für eine konkrete Bibliothek zum Sammelauftrag gehört, soll hier nicht weiter diskutiert werden. Offensichtlich ist aber eine enorme Ausweitung der Sammlungen möglich und im Einzelfall geboten. Wie immer die Grenze gezogen wird, muß die Begrenzung flexibel gehandhabt

werden: Allgemeine Überlegungen über die Bestandserschließung, Bereitstellung und Bestandssicherung dürfen sinnvollerweise keines dieser digitalen Medien ausschließen.

Die Organisation digitaler Materialien

Für die Digitale Bibliothek wirft diese Entwicklung spezifische Fragen der Organisation auf. Im Hintergrund der Digitalen Bibliothek stehen digitale Speicher, Rechner und Netze. Die digitalen Speicher sind in ihrem technischen Kern Speicher analoger Signale, die freilich anders als in der tradierten Analogtechnik beispielsweise des Tonbandes klar unterschieden werden können. Typischerweise sollen dazu nur zwei binäre Signale getrennt werden, etwas salopp *Null* und *Eins* genannt. Die Besonderheit digitaler Speicher liegt also in ihrer Beschränkung auf wenige Signale und in deren digitalen Prozessierbarkeit mittels Computern und ihren zugehörigen Programmen. Durch die Beschränkung auf klar unterscheidbare Binärsignale sind wesentlich bessere Fehlerentdeckungs- und Fehlerkorrekturverfahren einsetzbar als bei analog gespeicherten Objekten. Trotzdem ist die digitale Speicher- und Übertragungstechnik nicht „fehlerfrei,“ denn ihre materielle Basis bleibt auch mit allen Fehlerkorrekturtechniken eine analoge, fehlerbehaftete Signaltechnik. Im Alltag ist dies gelegentlich zu spüren und unter langfristigen Perspektiven keineswegs zu vernachlässigen. So garantieren optische Speichermedien wie die CD-ROM eine typische Fehlersicherheit von einem Fehler auf 10^{12} Speichervorgängen. Ein nichtkorrigierbarer Fehler tritt demnach im Schnitt bei jeder tausendsten nach dem Stand der Herstellungstechnik erzeugten CD-ROM auf; bei selbst gebrannten CD-ROMs ist dies deutlich häufiger. Diese an sich hohe Fehlertoleranz wird bezahlt mit einem erheblichen zusätzlichen Spei-

cheraufwand von etwa zwei Dritteln der gesamten Kapazität einer CD-ROM-Scheibe. Auch sind digitale Speicher inhärenten Materialfehlern oder schädigenden Umwelteinflüssen nicht entzogen. Zusicherungen der Speichermedienhersteller über die Lebensdauer ihrer Produkte, die von idealisierten Lager- und Nutzungsbedingungen ausgehen, müssen deshalb mit angemessener Vorsicht gesehen werden.

Davon unabhängig zeigen sich auch zwischen den Produkten unterschiedlicher Hersteller immer wieder deutliche Unterschiede in der inhärenten Fehleranfälligkeit – und gelegentlich sogar zwischen unterschiedlichen Fertigungspartien des gleichen Herstellers.

Digitale Backup Strategien

Anders als bei klassischem analogem Sammelgut ist es technisch vergleichsweise einfach von digitalen Objekten qualitativ gleichwertige technische Kopien herzustellen. Die Erstellung und Pflege dieser Kopien muß organisatorisch durch eine Backup-Strategie geregelt werden. Eine häufig anzutreffende Backup-Strategie für veränderliche Datenbestände besteht darin, drei „lebendige“ Varianten zu unterhalten: Die Arbeitskopie A, deren direktes Backup B1 sowie das zweitälteste Backup B2. In regelmäßigen Abständen wird von der aktuellen Arbeitskopie A ein neues Backup erstellt, indem immer abwechselnd B1 oder B2 überschrieben wird. Außerhalb des Kopiervorganges existieren also stets drei Generationen der Daten. Dies hat sich bei Arbeitsdaten, die häufig verändert werden als vorteilhaft herausgestellt. Da aber in Bibliothek und Archiv typischerweise unveränderliche digitale Objekte gesammelt werden, sind auch andere Strategien denkbar – und im Einsatz. Des Weiteren ist die Frage zu klären, wie weit vollständige Backups oder inkrementelle Backups sinnvoller sind

und ob eine lokale oder eine vernetzte Backupstrategie gewählt wird. Wichtig ist auch der Zeitpunkt und die Entscheidung über einen manuellen oder automatischen Start eines Backups.

Während Backups bei frei verfügbarem digitalem Sammelgut als vernünftige Verbesserung der Bestandsvorsorge unbedingt durchzuführen sind, kann dies bei fremd verfügbarem Material schon auf Grund der Lizenzbedingungen zum Problem werden. Insbesondere ist die Umgehung eines „wirksamen Kopierschutzes“ inzwischen urheberrechtlich untersagt, sofern keine besonderen Regelungen vereinbart sind: „Wirksame technische Maßnahmen zum Schutz eines nach diesem Gesetz geschützten Werkes oder eines anderen nach diesem Gesetz geschützten Schutzgegenstandes dürfen ohne Zustimmung des Rechtsinhabers nicht umgangen werden (§95a Abs.1 UrhG).“ Angesichts der problematischen Lebensdauer digitaler Speichermedien gefährdet diese Vorschrift den Bestand selbst dann, wenn die Sammelstelle zum Nachkauf defekter Medien bereit wäre, da kein Verlag eine langfristige Lieferbarkeit garantiert. Im Falle digitaler Artefakte gefährdet die Urhebergesetznovelle den gesellschaftlichen Sammelauftrag.

Zu beachten wäre auch, daß derartiges Material typischerweise der Produkthaftung durch den Hersteller unterliegt – wenn ein Fehler rechtzeitig entdeckt wird. Die neue urheberrechtliche Regelung darf auch unter dem Aspekt der Sammlung und langfristigen Bereitstellung nicht das letzte Wort sein.

Integrität und Authentizität

Bestandshaltung verlangt den unveränderten Erhalt des gespeicherten Sammelguts. So gehört es zukünftig zu den Aufgaben sammelnder Stellen, durch technische Maßnahmen sicherzustellen, dass illegale Manipulationen

an elektronisch veröffentlichten Publikationen nicht möglich sind. Zu sichern sind die Integrität gespeicherter Dokumente ebenso wie deren Urheberschaft (Authentizität), neben anderen Metadaten. Während dies bei analogem Druckmaterial mit hinreichender Sicherheit durch das technische Verfahren des Drucks und die Organisation der Archivierung gesichert ist, stellen sich bei digitalen Objekten neue Herausforderungen. Die spurlose Manipulierbarkeit eines digitalen Objektes erweist sich als problematisch, ebenso wie die vergleichsweise einfache Durchführung einer solchen Manipulation. Die Gefahr einer Manipulation zugänglicher digitaler Objekte scheint beim derzeitigen Stand der IT- und Netzsicherheit relativ hoch. Die Sicherung der Authentizität und der Integrität wird damit sehr wichtig. Verfahren der Authentifizierung und Integritätssicherung verursachen aber zusätzlichen Aufwand und, wenn dies gesetzlichen Vorgaben folgen soll, auch spürbare Kosten. Hier muß im konkreten Fall eine ausgewogene und angemessene Bestandssicherungsstrategie entwickelt werden.

Bei rechtlichen Verpflichtungen, wie sie etwa bei der Speicherung von Prüfungsarbeiten, Dissertationen oder Habilitationen bestehen, können digitale Dokumente signiert werden. Diese Signatur ist sogar gesetzlich geregelt („Signaturgesetz SigG“, Artikel 3 des Informations- und Kommunikationsdienste-Gesetz – IuKDG vom 11.6.1997). Naheliegenderweise steht der Bestandsschutz nicht im Zentrum dieser gesetzlichen Regelungen. Diese sollen vor allem finanzielle Transaktionen und staatliche Hoheitsakte regeln. Entsprechend wenig werden organisatorische und finanzielle Aspekte der Bibliotheken berücksichtigt.

Zur Handhabbarkeit der digitalen Signatur gibt es deshalb widersprechende Ansichten. So wird als problematischer Zug der digitalen Signatur als Teil

einer Bestandssicherungsstrategie der technisch begründete Zwang zu ihrer regelmäßigen Erneuerung gesehen. Diese scheint aber unvermeidlich wegen der begründeten Befürchtung, vertretbarer kryptografischer Schutz könne nach einigen Jahren wertlos werden. Das Dilemma wäre nur durch ein kostengünstiges und unaufwendiges Verfahren zu überwinden, wie z.B. eine gegenüber den gesetzlichen Regelungen einfachere Zertifizierung im Bibliotheksverbund – die freilich noch zu spezifizieren wäre.

Ein anderer Ausweg bestünde darin, dass Banken und andere Finanzinstitute auf breiter Basis Verfahren zur Signatursicherung einführen. Dies würde endlich zu einer billigeren und verfügbareren Technik führen, doch solange diese Organisationen sich nicht positiv zur digitalen Signatur bekennen, was sie vor allem aus Kostengründen, aber auch aus wettbewerbsstrategischen Überlegungen heraus scheuen, liegt zwar ein guter Ansatz vor, es fehlt aber an einer handhabbaren, kostengünstigen Durchsetzung.

III. Speichermedien und Speichermaterialien für multimediale Artefakte

Die langfristige Speicherung hängt wesentlich, wenngleich nicht ausschließlich, von den eingesetzten materiellen Trägermedien ab. Bei digitalen Medien besteht bislang keine seriöse Erfahrung über die langfristige Datenspeicherung auf einem festen Träger. Eine gewisse Ausnahme bilden Lochstreifen und Lochkarten, die ja bereits 1890 eingeführt wurden – freilich in einem heute vergessenen Format. Doch auch für die Weiterverarbeitung neuerer Lochkarten oder Lochstreifen werden entsprechende Karten- bzw. Streifenlesegeräte benötigt, deren Beschaffung und Wartung heute kaum noch möglich ist. Einzig von Menschen lesbare Ausdrücke auf Papier haben den informationstechnischen Wandel bislang überstanden.

Entsprechend gilt auch schon für viele Formen magnetischer und magneto-optischer Speichermedien, die in den letzten Jahrzehnten eingeführt und inzwischen technisch überholt wurden. Schon die stete Weiterentwicklung der Magnetbandkassetten lässt keine Hoffnung aufkommen, dass es jemals eine dauerhafte Archivierung auf festen Medien geben wird. Aus der Erfahrung von Versicherungen und Banken, die ja eine langfristige Datenhaltung über viele Jahrzehnte betreiben müssen (z.B. entlang der Laufzeit von Lebens- und Rentenversicherungsverträgen) folgt, daß nur eine dynamische Speicherung mit regelmäßigem Umkopieren auf neue aktuelle Datenträger für die längerfristige Archivierung zukunftssicher sein kann.

Speichermedien lassen sich in dauerhaft beschriebene („einmalig beschreibbare“) und in mehrfach beschreibbare Formen einteilen. Mehrfach beschreibbare optische Speichermedien sind flüchtiger als dauerhaft beschreibbare, doch auch die dauerhaft(er)en digitalen

Speichermedien wie CD oder DVD erweisen sich als nicht *sehr* dauerhaft. Eine Bestandssicherungsstrategie wird also entweder auf vergleichsweise kurzfristige Speicherung mit wieder beschreibbaren Medien wie etwa „Plattenfarmen“ und Bandkassettenarchive oder auf einer Folge von Umkopiervorgängen mit nur einmal beschreibbaren Medien beruhen müssen. *Multimediale Artefakte müssen dynamisch, in einer Folge von Umkopieren und gegebenenfalls Umkodierung gespeichert werden.*

Papierspeicher

Während Lochkarten zu den praktisch nicht mehr lesbaren Altbeständen der Datenverarbeitung gehören, bilden Papierausdrucke entgegen allen Phrasen vom „papierlosen Büro“ noch immer ein Speichergrundelement im Alltag. Mit den Fortschritten der Druckertechnik bis hin zum *Publishing-on-Demand* sind sie als Bestandssicherungsverfahren zu betrachten. Vorteilhaft ist das einfache Eingliedern in die papierenen Buch- und Zeitschriftenbestände. Deren Nachteile sind gleichfalls bekannt. Im Kontext der Digitalen Bibliothek sind dies insbesondere die schwierige Erschließung und Suche, sowie die vergleichsweise hohen Anforderungen an die räumliche Speicherung. Als erheblicher Nachteil erweist sich die mediale Beschränktheit des Papierspeichers: Audio, Video, interaktive Hypertexte oder Multimediaproduktionen können auf Papier nicht adäquat gespeichert werden. *Papierausdrucke sind deshalb für den Kernbereich multimedialer Datenbestände unbrauchbar.*

Mikrofilm und *Computer Output on Microfilm* – analoge Speicherlösungen

Mikrofilm verlangt Lagerungsbedingungen, die einigen Aufwand mit sich bringen, insbesondere als Farbmikrofilm. Die Verwaltung großer Filmmengen ist aufwendig. Sofern diese Bedingungen erfüllt sind, kann Mikrofilm ohne weitere Fürsorge über lange Zeiträume gelagert werden; man rechnet hier mit mindestens 100 Jahren, bevor erneutes Umkopieren notwendig wird.

Verfahren, mit denen digitale Daten direkt auf Mikrofilm ausgegeben werden können, sind bekannt (COM = Computer Output on Microfilm). Es gibt eine Reihe von Firmen, wie z.B. die Swisscom, die ihre langfristig zu haltenden Daten auf computerbeschriebenem Mikrofilm sichern. Es gibt Dokumentenmanagement-Systeme, die Computer Output on Microfilm integrieren, z.B. das System DoRIS der Firma Haessler, das u.a. im Schweizer Bundesverkehrsministerium eingesetzt wird. Moderne Mikrofilmplotter wie z.B. Microbox Polycom, können in ein Rechnernetz eingebunden und so konfiguriert werden, daß Primärinformationen und Metadaten gemeinsam so auf Film gesichert werden, wobei eine Redigitalisierung über Mikrofilmscanner möglich ist.

Die Investitionskosten für COM-Anlage und Filmmagazin sind freilich hoch, aber auf längere Sicht sind die Betriebskosten für die Filmlagerung und das Rückscannen niedriger als für regelmäßige Migrationszyklen und die Aufbewahrung digitaler Daten. Zu beachten ist freilich das Problem des technisch geregelten Zugriffs, womit es für Bibliotheken ausscheidet. Bei Archiven spielt dies aber eine untergeordnete Rolle. So berichten die Staatlichen Archive Bayerns, daß im Durchschnitt wird von den über 200

laufenden Kilometer an Beständen jährlich nur etwa ein Prozent benutzt werden.

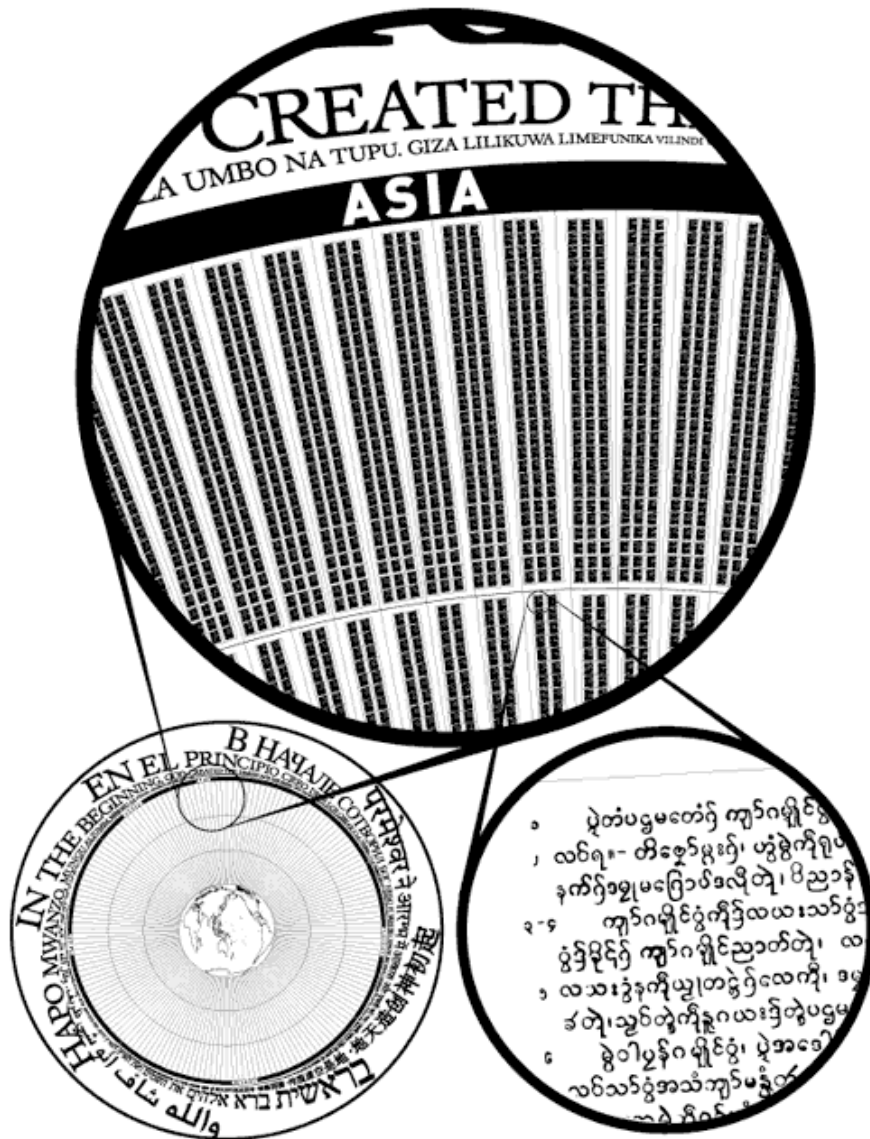
Mikrofilm verspricht unter guten Bedingungen eine fehlerfreie Lagerzeit von 100 Jahren. Viele Erfahrungen mit der Lagerung von Fotomaterial oder Kinofilm sind deutlich schlechter. Sie scheiden als langfristige multimediale Bildspeicher aus.

Mikrofilmspeicher sind eine spezielle Lösung für die Archivierung statischer Daten, also vor allem Texte, wenn auf die Vorteile digitalen Zugriffs und digitaler Erschließung verzichtet wird. Für multimediale Artefakte sind Mikrofilme nicht geeignet.

Rosetta Stone – eine HiTech-Speziallösung

Vor einigen Jahren wurde in den USA die *Long Now-Foundation* gegründet. Eines ihrer Ziele ist die Bewahrung wichtiger Zeugnisse menschlicher Kultur auf nahezu unzerstörbaren und jederzeit lesbaren Datenträgern. Die eingangs erwähnte Millenium-Uhr, die Dan Hillis konstruierte, gehört zu den Projekten der *Long Now-Foundation*. Auch eine „Bibliothek für zehntausend Jahre“ soll es geben. Als ein von der NSF und der Universität Stanford gefördertes Startprojekt wählte man die ersten drei Kapitel der Genesis, die in mehr als 1000 Sprachen in analoger Form auf einen äußerst langlebigen Datenträger geätzt werden, zusammen mit je 27 Textseiten mit Informationen zu jeder Sprache. Derzeit werden auf ihrer Webseite 1761 kodierte Sprachen ausgewiesen (<http://www.longnow.org> vom 10.8.2004).

Zu den werbewirksamen Aktionen der Long Now-Foundation gehört die Emission eines Datenträgers am 26.2.2004 mit Hilfe einer Ariane IV im Rahmen des Rosetta-Projekts der ESA, die im Jahr 2101 auf dem Kometen 67P/Churyumov-Gerasimenko landen soll.



Analogspeicher Rosetta Disk – links unten etwa in Originalgröße
(Quelle: www.rosettaproject.org am 10.8.04)

Der analoge Datenträger für das Projekt wurde von den Los Alamos Laboratories zusammen mit der Firma *Norsam* entwickelt. Die 3" große *Rosetta Disk* kann zwischen 30.000 und 350.000 Seiten Text oder Bilder mit einer optischen Verkleinerung um den Faktor 1000 speichern. Die Daten stammen von Mikrofilmen oder TIFF-Dateien und wurden mit einem Ionenstrahl in eine Nickellegierung geätzt.

Die Rosetta Disk soll gegen 2000 Jahre haltbar sein. Gelesen wird die Scheibe je nach Verkleinerung mit einem optischen Mikroskop oder einem Elektronenmikroskop, an das eine Digitalkamera und ein Rechner angeschlossen werden. Ein solches Lesegerät kostet derzeit um 10.000 US-Dollar.

Die analogen Verfahren des Papierausdrucks, der Mikrofilmausgabe oder der Rosetta Stone Technik mögen ihre Nische für die Backup-Sicherung analoger Bestände oder als Ergänzung zu digitalen Speichertechniken finden. Als Kernverfahren der Bestandsicherung in der Digitalen Bibliothek sind sie nicht ausreichend, schon deshalb, weil sie nicht alle digitalen Medienformate abbilden können. Diese analogen Speichertechniken bieten keine Sicherungsmöglichkeit für programmgestützte, interaktive oder dynamische Medien.

Optische Plattenspeicher

Als wesentlich kostengünstigere Dauerspeicher haben sich die optischen Speichermedien CD (Compact Disc) mit typischerweise 700 MB und DVD (Digital Versatile Disc) mit 4,5 GB bzw. 8,5 GB oder 17 GB bewährt – neben den Sonderformen der magneto-optischen Bauart (M-O).

Unterschieden wird die Bauform der im Presswerk hergestellten CD-ROM von der einmal beschreibbaren CD-R und der mehrfach beschreibbaren CD-RW. Ähnlich unterscheidet man DVD, DVD-R und DVD-RW, wobei es aus Wettbewerbsgründen auch noch die Bauformen DVD-R und DVD+R eingeführt wurden. DVD-RAMs sind eine spezielle Form der DVD-RW.

Eingeführte optische Medien sind vergleichsweise billig, robust und leicht handhabbar. Ihre Fehlerstabilität liegt bei etwa einem unkorrigierbaren Fehler bei 10^{13} bis 10^{15} gespeicherten Bits (CD-ROM bzw. DVD). Das entspricht einem unkorrigierbaren Fehler bei 1000 CD-ROMs. Diese Fehlerate ist aber alterungs- und umweltbedingt und sie ist bei selbst beschriebenen Medien (CD-R) und insbesondere bei wieder beschreibbaren Formen wie der CD-RW deutlich schlechter.

Ein offenes Problem bleibt die reale Haltbarkeit der optischen Medien. Sie wird bei guter Lagerung an Hand von Simulation und Tests auf ein oder zwei Jahrzehnte und mehr geschätzt. Da Erfahrungen seit 1982 bei der Audio-CD und seit etwa 1987 bei der CD-ROM vorliegen, mag man dies nachvollziehen – oder auch nicht. Problematischer als die Audi-CD und die in Massenaufgabe gepresste CD-ROM sind jedoch zweifellos die mehrfach beschreibbaren Bauformen wie die CD-RW oder die DVD-Ram.

Technologische Veralterung wird zum zentralen Problem auch bei den optischen Speichern. So wird die DVD, die gerade erst eingeführt wurde, soll demnächst durch Blu-Ray oder HD-DVD oder einer ähnlichen Variante mit einer vielfachen Speicherkapazität ergänzt und abgelöst. Binär kodierte *BluRay Recorder* wie der SONY BDZ-S77 mit 23 GB Kapazität sind seit April 2004 in Japan verfügbar; *HD-DVD-Recorder* mit ähnlichen Leistungsdaten sind angekündigt. Deren Nachfolger werden auch schon

genannt, z.B. aus den Londoner Forschungslabors des Imperial Colleges wird ab 2010 die nicht bloß binär, sondern mehrwertig kodierte *Multiple-plexed Optical Data Storage-Scheibe* (MODS) mit 250 GB – 1 TB angekündigt, mit der fast 500 Stunden Film gespeichert werden könnte.

Die besonderen Bauformen der wiederbeschreibbaren magneto-optischen Techniken werden in Nischenmärkte verdrängt, wo über ihre technologische Überlebenschance kaum etwas vorhergesagt werden kann. Ihre technologische Veralterung ist besonders problematisch, da mehrfach beschreibbare magneto-optische Speicher oft als sehr zuverlässige Backup-Speichermedien dargestellt werden – in gewisser Ignoranz bestätigter durchschnittlicher technologischer Lebensdauer.

Dennoch zählen optische Speichermedien, insbesondere in der einmal beschreibbaren Form als CD-ROM oder CD-R neben den Magnetplatten und Magnetbändern zum Kernbestand der multimedialen Speichermedien.

Die Iridium-CD - Datenspeicher

Bei der schon vor etlichen Jahren vorgeschlagenen Lösung werden mit einem der Rosetta-Disk ähnlichen Verfahren die digitalen Daten als analogisierten Bitstrom auf einen Iridium-Träger geätzt. Zudem werden die für eine Dekodierung des Bitstroms notwendigen Informationen als analoger Klartext in extremer Verkleinerung auf den Träger gebracht. Mit einem Elektronenrastermikroskop können diese Informationen gelesen werden. Ein Programmierer sollte dann in der Lage sein, anhand dieser Informationen einen *Viewer* zu programmieren, mit dem die Daten gelesen werden können. Praktische Einsätze dieser Technik in größerem Maß liegen bislang nicht vor.

Magnetische Plattenspeicher

Magnetische Plattenspeicher sind die dominante Technik des wiederbeschreibbaren Sekundärspeichers. Üblich sind fest verbaute Magnetplattenspeicher, es gibt daneben auch magnetische Wechselplatten, die in bestimmten Nischen, z.B. der Druckindustrie verbreitet sind. Sie werden freilich mehr und mehr durch optische Wechselmedien abgelöst.

Die im Markt übliche durchschnittliche Speicherkapazität wächst mit den technologischen Fortschritten ähnlich wie bei Halbleiterspeichern mit etwa konstanter Verdoppelungsrate von etwa 20 Monaten, d.h. in elf Jahren verhundertfacht sich die marktübliche Speicherkapazität einer Magnetplatte. Die derzeit üblichen Formfaktoren sind Platten mit 3,5" Durchmesser, ergänzt durch Platten zum Einsatz in Notebooks mit 2,5" Durchmesser. Für Kameras, MP3-Spieler und andere digitale Geräte werden auch Platten mit kleineren Durchmessern verbaut. Bei Schnittstellen hat sich IDE/ATA durchgesetzt, bei Servern gibt es auch noch das (teurere) SCSI und zunehmend Glasfaseranschlüsse für Hochleistungsserver. Typische Speichergrößen liegen über einhundert GB; einzelne Festplattenstapel werden bis zu einem halben Terabyte angeboten. Terabyte-Speicher sind in den nächsten zwei Jahren zu erwarten.

Magnetplattenspeicher arbeiten mit ausgereiften Fehlerkorrekturverfahren, die bei den derzeitigen Bauformen im normalen Betrieb Fehlerraten von einem nichtkorrigierbaren Fehler bei etwa 10^{15} Schreibvorgängen sichern sollen. Obwohl dies sehr niedrig ist, kommen doch immer wieder Plattenfehler bis hin zu Totalausfällen vor. Redundante Speicheraufbauten aus mehreren Platten (RAID-Speicher) könne diese Fehlerraten noch

einmal deutlich reduzieren. Manche RAID-Formen lassen den Austausch von einzelnen Platten sogar im laufenden Betrieb zu. Dennoch sind organisatorische Maßnahmen wie Backup-Strategien unerlässlich, um die letztlich unvermeidlichen technischen Speichersicherheitsprobleme zu meistern.

Magnetplatten sind die dominante Speichertechnik für mehrfach beschreibbare Medien und sie werden dies in den nächsten Jahren auch bleiben. Sie werden ergänzt durch optische Speicher und Magnetbänder für die archivierende Speicherung.

Magnetbänder

Magnetbänder gehören zu den ältesten Speichermedien. Ihre Verwendung ist heutzutage wegen der sequentiellen Beschränkung des Zugriffs auf Backup- und Archivierungszwecke im Serverbetrieb beschränkt. Es gibt eine Reihe von Formfaktoren, deren regelmäßiger technologischer Wandel ein regelmäßiges Umkopieren verlangt. Die Speicherkapazitäten orientieren sich an den Größen aktueller Festplatten.

Magnetbänder sind als Backup- und Archivspeicher für multimediale Artefakte gut geeignet. Ihre Verwendung für die Bereitstellung von Multimediaprogrammen wird durch die sequentielle Zugriffstechnik stark eingeschränkt.

Direktzugriffsspeicher - Halbleiterspeicher

Aus der technischen Entwicklung heraus scheint es schlüssig, daß Speicher mit mechanischen Zugriffstechniken, wie Bänder oder Platten, über

mittlere Frist von Speichern mit direktem Zugriff, also frei von mechanischem Zugriff, abgelöst werden. Die gängige Bauform sind Halbleiterspeicher: Schließlich wird die digitale Technologie wesentlich durch den Fortschritt der Halbleitertechnik bestimmt. Das Mooresche Gesetz, eine Art Generalplan der Halbleiterindustrie, verfolgt seit Jahrzehnten das bislang immer wieder erreichte Ziel, die produzierbare Kapazität typischer Halbleiterspeicherchips in etwa 18 Monaten zu verdoppeln. Dies führte alle zehn Jahre zu einer Verhundertfachung der typischen Speicherkapazität marktüblicher Rechner – seit nunmehr vier Jahrzehnten.

Bei Halbleiterspeichern werden derzeit dynamische und statische Bauformen unterschieden, wobei die Inhalte dynamischer RAM-Speicher in sehr kurzen Zeitintervallen durch eine Ansteuerschaltung „aufgefrischt“ werden müssen; statische RAM-Speicher brauchen diese Auffrischung nicht. Für die langfristige Speicherung sind weder dynamische noch statische RAM-Speicher geeignet, da sie ihre Daten bei Wegfall der äußeren Energieversorgung verlieren.

Flash-Speicher, wie sie in z.B. in digitalen Kameras als *Compact Flash*, *SmartMedia*, *Memory Cards* oder *Memory Sticks* eingesetzt werden. Sie speichern Daten mit einer Lebensdauer von zehn Jahren und mehr mit vielen tausenden Lese-Schreibvorgängen. Mit den Flash-Speichern technologisch eng verwandt ist die Bauform EEPROM (*Electrically Erasable Programmable Read Only Memory*), eine Variante der ROM-Speicher, die als langfristige Halbleiterspeicher geeignet sind.

Flash-Speicher sind wahlfrei auslesbar, aber nur in Blöcken beschreibbar, weswegen sie nicht als Ersatz von RAM-Speichern dienen können. Darüber hinaus werden Flash-Speicher aus ökonomischen Gründen bis jetzt nur in

speziellen Einsatzgebieten und in relativ kleinen Bauformen (unter acht GB) eingesetzt, die freilich schon die Größe einer einfachen DVD erreichen.

Eine neue von der IBM entwickelte Bauform, die jetzt erste kommerzielle Anbieter findet (*Cypress Semiconductors*), sind die Magnetischen RAM-Speicher (MRAM), die ihre Informationen ohne zusätzliche Energieversorgung halten. Über ihren praktischen Einsatz kann noch nichts ausgesagt werden.

Halbleiterspeicher wie Flash-Speicher oder MRAM besitzen das Potential optische Speicher als mehrfach beschreibbare Lesespeicher abzulösen. Ob und wann dies geschehen wird, hängt von ökonomischen Faktoren ab, aber auch von der Entwicklung der anderen Bauformen.

Holografische Speicher

Als viel versprechende Form optischer Speicher werden immer wieder holografische Speicher genannt, zu denen eine fleißige Forschergemeinde existiert, deren Arbeit aber noch nicht zu kommerziellen Formen geführt hat. Am Institut für Mineralogie der Universität zu Köln gelang es beispielsweise, digitale Daten als Hologramme in speziell für diesen Zweck gezüchtete LiNbO₃- oder LiTaO₃-Kristallen zu speichern. Dieser Speicher soll über 100 Jahre stabil sein. Mit gepulsten Lasern ist möglich, sehr große Datenmengen, wie z.B. einen digitalisierten 35mm-Spielfilm, in sehr kurzer Zeit zu speichern und so rasch wieder auszulesen, dass ein Video ruckelfrei angesehen werden kann. Als Übertragungsrate wird ein MB innerhalb einer µs angegeben. Den Prototypen eines Lesegeräts gibt es als Labormuster, aber eine kommerzielle Lösung wird noch nicht angeboten.

Auch Firmen wie z.B. der Bell Lab-Offspring *InPhase Technologies* in Denver, Colorado hat ein holografisches Speichersystem vorgestellt, das auf der Fläche einer CD 88 Stunden Video oder 15 Stunden HDTV-Video speichern kann (das wären ca. 150-200 GB). Die Einführung der holografischen *Tapestry*-Speicher wurde einst für 2004 angekündigt – ein Termin der nun auf 2006 verschoben ist.

Holografische Speicher können als Direktzugriffsspeicher entwickelt werden. Sie sind zwar optische Speichermedien, können aber wie Halbleiterkristalle frei von mechanischen Zugriffen gebaut werden.

Ob und in welchem Umfang sich holografische Speicherbausteine durchsetzen und zu welchen Kosten, bleibt vorerst abzuwarten.

Das Netz als Speicher

Funktional lassen sich die bibliothekarischen Aufgaben der Sammlung, Bestandssicherung und Bestandsverfügbarkeit von der reinen digitalen Speicheraufgabe abtrennen. Dann wäre das Netz (das Internet) der Speicher - realisierbar von spezialisierten Auftragsnehmern, die beispielsweise ein SAN (*Storage Array Network*) betreiben. Eine solche *Outsourcing*-Lösung würde freilich eine *Neuorganisation* von Sammlung, Bibliothek oder Archiv bedeuten. Es wäre aber auch eine Chance der Bibliothek, die erst durch die Digitalisierung entsteht. Gerade weil die digitalen Speichermedien keinen direkten Bezug mehr zum gespeicherten Inhalt haben, würden sich die Aufgaben der Sammlung, der inhaltlichen Erschließung und der Bereitstellung des Zugangs völlig von der Speicherung trennen lassen. Dies ist bei einigen digitalen Dokumentenservern schon angelegt. In einem Universitätsbetrieb könnte dies relativ einfach als Aufgabenteilung zwischen

Bibliothek und Rechenzentrum realisiert werden. Selbstverständlich sind auch *Outsourcing*-Lösungen unter anderen bibliothekarischen Bedingungen denkbar.

Festzuhalten bleibt: Regelmäßiges Umkopieren ist somit eine notwendige Forderung zur Bestandserhaltung. Magnetplatten und optische Speicher stehen im Kern der Speichertechnik, Magnetbänder bleiben im Serverbereich Backup- und Archivmedien der Wahl. Externe Halbleiterspeicher (Flash Speicher) besitzen ein großes Potential als externe Speicher auf mittlere Frist. Mit welchen Technologien und wo dies stattfindet, ist nicht mit letzter Sicherheit zu klären.

Leider bietet Umkopieren keine hinreichende Bestandserhaltungsstrategie. Trotz der recht beschränkten Lebenszeit technischer Digitalspeicher mit der Folge unlesbarer Medien scheint die Wahrscheinlichkeit größer, daß Daten durch technologische Alterung, also veraltete Geräte und Software, verloren gehen werden.

Zugriffsvermittlung über elektronische Netze

Besonderheiten entstehen bei der Datenbereitstellung über Netzzugriff als Bibliotheksangebot, wie es von Zeitschriftenverlagen seit einiger Zeit angeboten wird. Die zeitlich beschränkte Zugangserlaubnis zu elektronischen Zeitschriften, die auf einem zentralen Server unter Verlagshoheit stehen, wandelt die Bibliothek zu einem Makler, der u.U. nicht einmal mehr die Eingriffsmöglichkeiten besitzt, die fremd erzeugte Dokumente bieten (soweit dies nicht durch gesetzliche oder vertragliche Schranken verboten ist). Unter Gesichtspunkten langfristiger Datensicherung oder gar

Archivierung kann dies unerwünschte Auswirkungen haben. Eine bloße Zugriffsregelung ist mit einer Bestandssicherung nicht mehr vereinbar.

Aus sachlichen Gründen ist dringend darauf zu achten, daß die zu sichernden und zu bewahrenden digitalen Objekte dauerhaft und langfristig in die Verfügungsgewalt der sammelnden Stelle übergehen. Dies muß freilich nicht zwingend mit einer Präsenz der Daten in der Bibliothek verbunden sein. Es ist denkbar, daß Daten bei einer vertraglich gebundenen, vertrauenswürdigen Stelle hinterlegt und im Netz verfügbar sind – unabhängig von der wirtschaftlichen Zukunft des Verlags, gesichert durch einen dauerhaften Vertrag bis zum Ablauf eventueller Urheberrechte. Danach sollen sie in die freie, nicht-exklusive Verfügung der Bibliothek übergehen.

IV. Technische Speicherformate – vom Text zum Multimediaobjekt

Im Digitalen sind technische Formate, logische Formate und algorithmisierte Formate zu unterscheiden.

Die historisch gewachsen wichtigste Dokumentform ist der Text, typischerweise in Form eines Buches oder einer Zeitschrift. Obwohl Texte meist hochgradig strukturiert sind, werden sie aus Sicht der Informatik meist als „frei formatierte“ Dokumente betrachtet und verwaltet. Dies erlaubt große Freiheiten beim Erzeugungsprozeß, stellt sich aber mehr und mehr als Hindernis bei Speicherung, Präsentation und gezieltem Zugriff, der „Suche“, dar. Einen Ausweg aus dieser Einschränkung bei selbst bestimmt erstellten Dokumenten bietet die Zufügung von *Metadaten* (bei Standards wie SGML, XML; in beschränktem Maße *Style Sheets*). Fremd erzeugte Dokumenten mit Metadaten können in gleicher Weise behandelt werden, sofern eine hinreichende Übereinkunft zu den Metadaten getroffen wurde.

Grundsätzlich ist auf eine sparsame Verwendung unterschiedlicher Formate zu achten. Offen gelegte Formate sind gegenüber proprietären, einem Geschäftsgeheimnis unterliegenden Formaten (wie z.B. Word .doc) unbedingt zu bevorzugen, standardisierte Formate sind gegenüber firmenabhängigen zu bevorzugen. Obgleich die tatsächliche Verbreitung eines Formats natürlich von Gewicht ist, sollten zentrale Leitfragen sein:

- Wird dieses Format in einem in der Zukunft (in einem Jahrzehnt oder in mehreren Jahrzehnten) noch unterstützt?

- Ist ein Austauschformat verfügbar, so daß das Objekt mit unterschiedlichen Programmen präsentiert und umkopiert werden kann?
- Gibt es mehr als nur eine unterstützende Firma?
- Wie sicher werden die Firma oder Rechteinhaber zukünftig existieren?
- Ist die Nutzung durch Lizenzen eingeschränkt oder frei verfügbar?
- Ist das Format offen gelegt, so dass es nachträglich umkodiert werden kann?
- Ist das Format firmenübergreifend standardisiert?

Häufig ist eine frühzeitige Umformatierung sinnvoll, um später zusätzliche Altlasten zu vermeiden.

Leider gibt es keinen entwickelten Kanon multimedialer Standard, der alle medialen Formen digitaler Objekte beschreibt. Standards müssen für jede Medienform identifiziert und bewertet werden. Eine erster Ansatz zur Erfassung unterschiedlicher Medientypen kann in der Registrierung der *Multipurpose Internet Mail Extensions* für E-Mail-Protokolle (der MIME-Typen gemäß Requests for Comment RFC2045 und RFC2046) durch die *Internet Assigned Names Authority* IANA vom November 1996 gesehen werden, der für die Zwecke der Langzeitarchivierung freilich nicht ausreicht.

Ein neuerer Ansatz, der über die bloße Registrierung der Datentypen hinaus geht, folgt mit dem Aufbau eines *Global Digital Format Registry* (GDFR) zu sehen, wie es von der DSpace User Group des MIT, der Harvard LDI Group u.a. vorgeschlagen wird (Abrams & Seaman, 2003), einer Entwicklung,

hinter der inzwischen eine bedeutende Gruppe internationaler Institutionen steht. Die Ziele des Formatregisters sind: „*The registry will maintain persistent, unambiguous bindings between public identifiers for digital formats and representation information for those formats.*“ Das Register wird als zuverlässige, treuhänderische Stelle zur langfristigen Pflege der Formatdefinitionen gesehen.

Texte

Texte werden als kodierte Bitketten gespeichert. Die Umsetzung der Codes (ASCII, Unicode) geschieht an Hand von Tabellen in programmierter Weise. Dies kann als zukunftssicher angesehen werden. Proprietäre Formate und Rechte sind zu beachten und nach Möglichkeit zu vermeiden. Zukunftssicher sind nur offene lizenzfreie und frei lizenzierte Formate. Problematisch bleibt der Präsentation eines Textes jenseits der ASCII-Struktur, etwa die Verwendung von Schriften und Satzanweisungen, deren langfristiger Einsatz keineswegs in gleicher Weise gesichert ist (s.u.).

Daten, Datenbanken, Tabellenkalkulationsprogramme

Formatierte Daten sind in ihrem Grundgehalt in Standardformaten wie ASCII oder BCD ablegbar. Typischerweise sind sie freilich mit Metadaten einer Datenbank oder eines Tabellenkalkulationsprogrammes versehen. Dafür gibt es spezielle Austauschformate wie z.B. CMV oder SYLK für Tabellenkalkulationsprogramme. Es bleiben aber trotzdem Forderungen nach zukunftssicherer spezifischer Hard- und Software.

Bei Datenbanken hat sich nach einer langen Geschichte unterschiedlicher Formate als Austauschformat inzwischen XML (*eXtensible Markup*

Language), ein offener Standard, durchgesetzt. XML ist eine vereinfachte Teilmenge von SGML und steht in lockerer Beziehung zur Hypertext Markup Language HTML, die ebenfalls aus SGML abgeleitet ist (technisch gesehen kann HTML als SGML-DTD interpretiert werden). Mit der „Extensible HyperText Markup Language“ (XHTML) ist der Übergang von HTML zu XML als Definitionsbasis vollzogen. XML wird wie SGML oder HTML als ASCII- bzw. Unicode-Text gespeichert und mit Hilfe eines Interpreterprogramms (*Parser*) grafisch präsentiert.

Die XML-Syntax ist einfacher als SGML-Beschreibungen. Entsprechend einfacher sind passende Parser zu erzeugen. Viele Web-Browser wie MS Internet Explorer, Netscape Navigator, Firefox und andere Mozilla-Derivate, Opera oder Safari können XML-Dateien direkt mit Hilfe eines eingebauten XML-Parsers darstellen (wenngleich in unterschiedlichem Umfang und mit unterschiedlicher Zuverlässigkeit). Dies geschieht in Verbindung mit *Stylesheets*.

Die Namen der einzelnen Strukturelemente für eine bestimmte Auszeichnungssprache lassen sich frei wählen. Auszeichnungssprachen können dabei unterschiedlichste Daten beschreiben, als wichtigstes Beispiel Text, aber auch Grafiken oder andere mediale Formen. Ein Grundgedanke von XML ist es, Daten von ihren Repräsentation zu trennen, also beispielsweise geostatistische Daten einmal als Tabelle und zum anderen als Karte auszugeben, wobei beide Anwendungen auf die gleiche Datenbasis im XML-Format zugreifen. Soll XML für den Datenaustausch verwendet werden, so werden die Datentypen in einer Dokumenttypdefinition (DTD) oder einem XML Schema festgelegt. XML ist nicht nur für Datenbanken von Bedeutung, es kann auch als zukunftsicheres Format für Texte und anderes Schriftgut angesehen werden. Dies bedeutet jedoch keineswegs,

daß für bereichsspezifische Datenbanken nicht auch andere Austauschformate verwendet werden (wie z.B. bei Bibliothekskatalogdaten MAB und MARC).

Programme

Computerprogramme sind einer doppelten Notation unterworfen, dem Quellcode und dem Maschinen- oder Binärkode. Der Quellcode beschreibt die Funktion eines Programms in einer für Menschen zumindest im Prinzip verständlichen Form, nämlich als Text in einer wohl definierten Programmiersprache. Der Binärkode beschreibt das maschinenausführbare aus dem Quellcode übersetzte Programm als Steueranweisung für ein bestimmtes Gerät und ein bestimmtes Betriebssystem. Bestandssicherung sollte sinnvollerweise auf beide Formen, den Quelltext ebenso wie die ausführbare Binärdatei zurückgreifen können, um gegebenenfalls den Quelltext an ein technologisch aktuelles Gerät und Betriebssystem anpassen zu können (*Umkodierung*). Falls eine Option besteht, sind deshalb *Open Source Programme* als Sammelgut ganz klar zu bevorzugen.

Für Programme, die nur als Binärdatei vorliegen, ist eine Vorhersage über ihre technologische Haltbarkeit schwierig. Ein denkbarer, aber bislang nur in Spezialfällen beschrittener Pfad der Migration könnte mit Hilfe von *Emulatoren* verfolgt werden. Positive Beispiele gibt es im Bereich der Hobby-Computerspiele, bei der musealen Simulation alter Computerhardware und -software, aber auch bei einigen kommerziellen Emulatoren wie der Java Virtual Machine oder den Apple 68K-Emulatoren für die PowerPC-Architektur.

Formatierte Texte

MS Word

Microsoft Word verwendet das proprietäre Speicherformat .doc. Da dies nicht offen gelegt wird, ist es als dauerhaftes Speicherformat von der Weiterentwicklung von Word durch die Firma Microsoft abhängig. Es gibt zwar eine Reihe von Office-Paketen, die das Word-Format importieren können, eine getreue Übertragung ist jedoch nicht gewährleistet und wird bislang auch nicht erreicht.

RTF

Das *Rich Text Format* RTF wurde von Microsoft als Austauschformat für Textverarbeitungsprogramme, insbesondere für MS Word entwickelt. Es ist von nahezu allen Textverarbeitungsprogrammen adaptiert worden. Eine volle Kompatibilität besteht allerdings nicht einmal zwischen unterschiedlichen MS Word-Versionen. RTF kann nicht als Standard für die langfristige Bestandserhaltung angesehen werden, obwohl es geeigneter ist als das nicht einmal offen gelegte .doc-Format.

TeX

TeX ist eine voll dokumentierte Dokumentenstrukturierungssprache und Satzanweisungssprache. Das Copyright liegt bei D.E.Knuth von der University of Stanford, der die Nutzung des Quellkodes frei gab (die Nutzung des Namens aber unter Konformitätsbedingungen stellt). TeX ist äußerst beliebt bei Mathematikern und Naturwissenschaftlern, da es viele Besonderheiten des mathematischen Satzes beherrscht. Der Austausch mit anderen Textformaten ist jedoch nicht einfach und die Grundkonzeption

mit einer strikten Trennung von Texterstellung und Satz verhindert seine Verbreitung außerhalb der genannten Bereiche, da die grafisch orientierte WYSIWYG-Darstellung („What you see, is what you get“) von vielen Nutzern bevorzugt wird. Als Archivformat steht TeX deshalb etwas isoliert dar.

PDF

Das *Portable Document Format* PDF ist aus der Seitenbeschreibungssprache Postscript entstanden. Obwohl PDF ein proprietär verwaltetes Format der Firma Adobe ist, ist die Datenstruktur offen gelegt, so daß andere Erzeuger- und Viewerprogramme, auch solche mit quelloffener Software, existieren. PDF ist nicht auf Texte und Formatierungen beschränkt; es kann ebenso Grafiken, Fotos oder Hyperlinks verwalten. Eine besondere Rolle spielt es als Druckausgabeformat bzw. auch zur Erzeugung von Bildschirmausgaben (*Display Postscript*).

Wegen seiner großen Bedeutung im Druck- und Satzgewerbe, ebenso wie der Tatsache, daß sehr viele Computerdrucker PDF-Interpreter enthalten, kann PDF trotz seiner proprietären Verankerung mit Einschränkungen als akzeptables Speicherformat angesehen werden.

HTML, SGML, XML

SGML wäre zur Bestandserhaltung sehr geeignet, wenn hinreichend viele Dokumente in SGML erstellt würden. Die Situation ist insofern ähnlich wie bei TeX. Es gibt freilich Hilfsmittel, um aus Worddokumenten SGML zu erzeugen. Das aus SGML abgeleitete XML gewinnt dagegen rasch an Bedeutung und Verbreitung, so dass dies eine zukunftsichere Notation zu

werden scheint. Positiv ist in beiden Fällen, dass die interne Notation in Unicode mit geeigneten *Document Type Definitions* erfolgt. HTML ist als reine Struktursprache ohne hinreichende Angaben zur Präsentation nicht zur dauerhaften Speicherung geeignet.

Schriften

Im einfachsten Fall liegen Schriften als Bitmaps fester Größe, also als grafische Notation evt. ergänzt um Farbangaben vor. Dieser Fall verliert an Bedeutung. Moderne digitale Schriften sind *Programme*, die in Postscript geschrieben sind und als Container von Zeichenumrissen, *Kerning*-Tabellen und gegebenenfalls weiteren Metadaten und Schrift-renderprogrammen eingesetzt werden. Neben *Postscript* werden *Truetype* und zunehmend *Opentype* eingesetzt. Die Rechte für diese Schriftbeschreibungsprogramme liegen bei Adobe, Microsoft und Apple sowie weiteren Lizenznehmern. *Dies ist eine problematische Situation, die derzeit nicht umgangen werden kann.*

Grafik, CAD

CAD-Systeme sind Vektorgrafiksysteme, deren Daten langfristig zugreifbar und austauschbar sein müssen. Vektorgrafiken beschreiben Kurvenstücke zwischen markanten Punkten und einfachen geometrischen Formen. Während die Vektoren als statische Beschreibung gespeichert werden können, verlangt ihre Präsentation programmierte Instruktionen, die von (auch zukünftig) verfügbarer Hard- und Software abhängen.

Ein offenes Austauschformat ist das vom amerikanischen *National Institute of Standards* NIST, vom *American National Standard Institute* ANSI und der *International Standards Organization* ISO standardisierte *Computer*

Graphics Metafile CGM-Format. Ein weiteres vom NIST entwickeltes Austauschformat ist die *Initial Graphics Exchange Specification* IGES. IGES definiert den Austausch zwischen 2D- und 3D-CAD-Daten. CGM und IGES sind als bestandswahrende Formate für diesen Spezialfall gut geeignet. Als neuere Entwicklung hat Autodesk, einer der führenden CAD-Hersteller, DesignXML –eine XML Ausprägung– als Austauschformat vorgestellt.

Neben CAD spielen Vektorgrafiken im Computergrafik und –designbereich eine wichtige Rolle. Hier ist Postscript in der Form *des Encapsulated Postscript* EPS das führende Format. Die Rechte liegen bei Adobe (wie bei Postscript selber), aber EPS ist mit allen Vektorgrafikprogrammen verträglich.

Flash ist ein beliebtes proprietäres Format für grafische Animationen. Die Rechte liegen bei Macromedia. Adobe hat (wohl nicht zuletzt unter dem Eindruck der in diesem Bereich marktbeherrschenden Stellung von Macromedia) ein offenes Konzept der *Scalable Vector Graphics* SVG vorgeschlagen und erzeugende Programme dafür vorgestellt. SVG ist unter dem Aspekt nicht-prorietärer, offener Standards zu bevorzugen: Die weite Verbreitung von Flash-Grafiken steht dem freilich entgegen. Die Akzeptanz von SVG ist dagegen bislang verhalten.

Fotos, Bilder

Digitale Bilder werden in einer Vielzahl von Formaten verwendet, die prinzipiell in die beiden Gruppen „Pixelbeschreibungen“ (Bitbilder, evt. in mehreren Ebenen) und „Vektorbeschreibungen“ eingeteilt werden können. Plotter können Vektorgrafiken direkt zeichnen. Ansonsten werden solche Vektorgrafiken in Bitmaps gewandelt („gerastert“), bevor sie ausgegeben werden.

Verfahren der Bildkompression sind algorithmische Formate, die Computerprogrammen nutzen. Dies setzt geeignete Hardware und Software voraus. Hinzu mögen rechtliche Beschränkungen kommen (z.B. LZW als, freilich auslaufender, patentierter Kern von GIF).

Bei Bitmaps haben sich umkomprimierte Formate wie TIFF (*Tagged Image File Format*) oder komprimierte Formate wie JPEG, JPEG 2000 und in geringerem Maße PNG durchgesetzt. Sie unterstellen freilich eine zukunftssichere Programmierung zur Dekodierung (und Kodierung).

Das Copyright für TIFF liegt inzwischen bei der Firma Adobe; ein LZW-Kern ist wie bei GIF von Unisys patentiert. Auch Xerox scheint Ansprüche auf TIFF-Teile zu erheben.⁴

Im September 2004 hat Adobe zur Nutzung in Digitalkameras die proprietäre, aber lizenzfreie „Digital Negative Specification“ DNG vorgeschlagen, die eine einheitliche umkomprimierte „Rohdaten“-Speicherung auf Basis von TIFF als Archivbildformat mit Metadaten anbietet. Adobe bietet dazu kostenlos nutzbare Konverterprogramme für unterschiedliche Kameratypen an.

JPEG ist zwar ein Ergebnis einer internationalen Standardisierung eines SO/IEC and ITU-T Komitees SC29, es kommt aber auch hier immer wieder zu Patentstreitigkeiten über einzelne Aspekte des Verfahrens. 2006 scheinen die letzten umstrittenen Ansprüche allerdings auszulaufen.

Nach Erklärungen des Standardisierungskomitees sind die „baseline“-Verfahren von JPEG und JPEG 2000 lizenzfrei: „SC 29 affirms and supports ISO policy that requires disclosure of the existence of Intellectual Property

⁴ <http://www.ietf.org/ietf/IPR/XEROX-RFC2301>

(IP) rights or pending rights (such as patents or pending patent applications), hereafter referred to as *IP rights*, associated with any technology submitted to SC 29/WGs for consideration for inclusion in any ISO/IEC standard. Specifically, SC 29 affirms the ISO policy of only considering technology that is free of *IP rights* or which is available on a royalty and license fee free basis or which is available under reasonable terms and conditions on a non-discriminatory basis.“ Es bleibt ein gewisses Restrisiko, daß unter den jetzigen Bedingungen wohl hingenommen werden kann.

Sowohl das *Portable Network Graphics* Format PNG wie das daraus abgeleitete JNG haben dagegen den Vorteil, quelloffene Standards zu sein, die bislang allerdings keine allzu große Verbreitung gefunden haben.

Video

Bei Bewegtbildern ist eine unkomprimierte Speicherung wegen des enormen Speicherplatzbedarfs kaum realisierbar und auch nicht sinnvoll. Es gibt typischerweise eine deutliche Trennung des Ausgangsmaterials vom distribuierten Dokument. Die Präsentation gespeicherter Bewegtbilddokumente stellt also spezifische Anforderungen an zukunftssichere Hard- und Software, die ohne standardisierte Austauschformate keine offene Langfristspeicherung erwarten lässt. Neben vielen spezifischen aufgabenorientierten Kompressions- und Speicherverfahren breiten sich die MPEG-Verfahren aus. Da MPEG-2 ein Standard der DVD und des digitalen Fernsehens ist, besteht eine gewisse Hoffnung, dass dieses Format noch einige Zeit gepflegt wird. MPEG-4 zeichnet sich als kommender Standard ab, der allerdings auch schon in Varianten verbreitet ist wie DivX;-), und DivX. XviD und OpenDivX sind verwandte quelloffene Implementierung

des MPEG-4 Standards. Am Horizont erkennbar werden weitere Standards wie MPEG-7 und MPEG-21.

Als Alternative gewinnt H.264 an Bedeutung, ein Standard, der ursprünglich für Videotelefonie und Videokonferenzen geplant wurde, aber auch Bewegtbilddarstellungen im Netz unterstützt und nun auch von seiner Qualität her zu einer Konkurrenz für MPEG-4 wird.

Ein wesentliches Problem entsteht beim Umkodieren komprimierter Daten, da dies zwingendermaßen zu einer Verschlechterung der Datenqualität führt (ähnlich der optischen oder akustischen Analogkopie).

Audio

Musiknotationen

Die Notation von Musik mittels Notenschrift ist im Computer derzeit nicht auf einen Standardcode zurückführbar. Es gibt aber verschiedene proprietäre Lösungen, die Marktakzeptanz haben. Eine geeignete XML-Notation wie MusiXML mag hier Abhilfe schaffen.

Eine andere instrumentennähere Notation ist MIDI (Musical Instruments Digital Interface) und seine Weiterentwicklungen. MIDI ist als Steuersprache (oder Programmiersprache) für elektronisch ansteuerbare Instrumente entworfen worden. Da Steueranweisungen in reiner ASCII-Notation erfolgen, kann MIDI wie ein Text gespeichert werden. Dies gilt allerdings nicht für „Samples“, musikalische Klangaufnahmen, die als Daten, nämlich als Tonaufnahmen gespeichert werden und in MIDI integriert werden.

Tonaufnahmen

Tonaufnahmen, ob analog oder digital, unterliegen immer Beschränkungen bei der Aufnahme und Speicherung. Die triviale Festlegung verfolgt ein vergleichsweise unspezifisches Hörmodell: eine feste Anzahl von Tonspuren, ein festgelegter Tonumfang, ein festgelegter Dynamikumfang. Als Standard hat sich AIFF (*Audio Interchange File Format*), die Tonspur der Audio-CD durchgesetzt, die zwei unkomprimierte Datenspuren enthält (Stereo, 20Hz-20KHz, 98 dB Signal-Rauschabstand).

Neben AIFF, dem Standardformat der Audio-CD haben sich inzwischen mehrere mehrkanalige „Heimkinoformate“ für die Audiospur der Film-DVD etabliert. Als wesentliche Kontrahenten stehen sich die Formate der Dolby Laboratories, darunter Dolby Surround, Dolby Pro Logic mit 5+1 bzw. 6+1 Kanälen und die Digital Theatre System DTS-Formate mit einer Serie von Varianten (DTS-ES, DTS NEO:6, DTS 96/24, DTS interactive, DTS HD/Lossles) gegenüber. Obwohl sie im DVD-Filmbereich die übliche Tonspur bilden, sind sowohl die Dolby- wie die DTS-Formate proprietär.

Zudem wird AIFF ergänzt durch unkomprimierte Audioformate höherer Qualität wie der Super Audio CD (SACD) mit einer 1-bit Abtastung bei 2,8 MHz und der DVD Audio mit monauraler, zwei- oder mehrkanaliger Abtastung mit einer bis zu 24 bit Abtastung bei 192 kHz. Zwischen beiden Formaten herrscht ein sogenannter „Krieg der Formate“, wobei beide derzeit wenig Marktdurchdringung erreichen.

Kompressionsverfahren verlangen eine programmierte Verarbeitung und stellen damit Anforderungen an zukünftige Hard- und Software. In Fällen wie MP3 (korrekter MPEG-1 Layer III) oder AAC (Advanced Audio Compression; Teil von MPEG-2 und MPEG-4) greifen auch noch

lizenzrechtliche Anforderungen. Die MP3-Kodierungstechnik mit einer Vielfalt möglicher Parametereinstellungen hat innerhalb eines Jahrzehnts eine enorme Verbreitung gefunden und AIFF im Computer- und Multimediabereich überholt. MP3 ist mit in der kommerziellen Verwendung mit Lizenzen belegt; hier etabliert sich parallel dazu die patentfreie Entwicklung .ogg als Tonspur des Multimediacontainers Ogg.Vorbis. Aus Gründen technischer Überlegenheit (höherer Kompressionsrate bei gleicher Qualität) scheint AAC jetzt MP3 abzulösen.

Aber selbst „freie“ und „verlustfreie“ Kompressionsverfahren wie FLAC stellen spezifische Hard- und Softwareanforderungen. Freilich gibt es patentfreie Quasistandards der Kompression wie Lauflängenkodierung und Huffman-Kodes.

In gewisser Weise kann man davon sprechen, dass AIFF das erste Jahrzehnt digitaler Tonaufspeicherung (auf der AudioCD) definiert hat und MP3 die vorherrschende Technik des zweiten Jahrzehnts geworden ist – beide ergänzt von vielen weiteren Formaten. Ob AAC nun das angebrochene dritte Jahrzehnt definiert, ist offen. Für die Langzeitarchivierung könnte dies der Beginn einer äußerst unangenehme Entwicklung werden: Müssen innerhalb eines Jahrhunderts 8-10 wesentlich verschiedene Basisformate verwaltet werden?

Gesprochene Sprache

Zur Speicherung von Sprache sind gegenüber der Speicherung von Musik gegebenenfalls stärkere Qualitätseinschränkungen möglich. Spezifische Standards sind aus den Musikspeicherstandards abzuleiten.

Kompressionstechniken

Digitalisierte Objekte können komprimiert gespeichert werden. Kompressionsverfahren werden naheliegenderweise meist nach ihrer Kompressionsleistung und nach ihrem Zeitverhalten beurteilt. Dies sind unter dem Aspekt langfristiger Speicherung oder Archivierung nachgeordnete Kriterien. Hier sind unter dem Aspekt langfristiger Nutzbarkeit quelloffene Verfahren wie z.B. (verlustfreies) GZIP oder BZIP2 gegenüber proprietären Verfahren zu bevorzugen. Speziell für Audiodaten liegen das quelloffene FLAC und andere, ähnliche Verfahren vor. Verlustbehaftete Verfahren werden typischerweise für einen einzelnen Einsatz bewertet. Sie sind jedoch ebenso auf ihre Einsetzbarkeit unter dem langfristigen Aspekt mehrfacher Umkodierung zu betrachten – womit sie streng genommen ausscheiden, da jede Umkodierung verlustbehaft kodierter Daten deren Qualität mindert.

Spiele und anderes interaktives Material

Computerspiele werden meist hochgradig proprietär hergestellt, um vorhandene Hardware bis an die Grenzen ihrer Leistungsfähigkeit auszunutzen. Es gibt aber Ansätze zur Emulation älterer Hardware und Betriebssysteme, beispielsweise für Commodore C64, Atari ST oder Amiga, um „technologisch verlorene“ Spiele auf aktuellen Geräten zu reaktivieren (Wintel, MacOS X). Ein bekanntes Beispiel ist die Open Source-Bewegung um MAME, den *Multiple Arcade Machine Emulator*. Das MAME-Team um Nicola Salmoria besteht aus etwa 100 Hobbyprogrammierern, die es bislang zur Version 0.85 des MAME-Emulators unter Windows XP gebracht haben.

Diese erfolgreiche Emulationstechnik kann als Vorbild für Emulationsprogramme zur Bestandssicherung multimedialer Daten dienen.

Multimedia

Multimediale Dokumente integrieren unterschiedliche digitale Datenformate von Text bis Bewegtbild mit Hilfe von Computerprogrammen. Durch die enge Verbindung von Daten mit Geräten und Programmen entstehen Dokumente, die mit Hilfe von Programmen und Geräten interaktiv und navigierbar genutzt werden können. Daraus folgen gleichzeitig sehr komplexe Anforderungen an eine zukunftssicherere langfristige Speicherung und Archivierung, da eben nicht nur statisch lesbare Dokumente vorliegen, sondern programmtechnische und gerätetechnische Anforderungen zur Nutzung des Multimediadokumentes zu erfüllen sind. Hinzu kommt die Notwendigkeit komprimierte Daten wie sie bei Ton-, Bild- und Bewegtbildanwendungen typisch sind, mit Hilfe von Programmen und Geräten zu dekodieren. Multimediale Dokumente sind also stets zusammen mit ihren programm- und gerätetechnischen Voraussetzungen zu behandeln.

Ein verbreitetes einfaches Multimediasystem ist das Präsentationsprogramm *Microsoft Powerpoint*. Obwohl das proprietäre Ausgabeformat .ppt von vielen anderen Präsentationsprogrammen gelesen werden kann, ist eine verlustfreie Übertragung einer multimediareichen Powerpoint-Präsentation praktisch ebenso schwierig wie die Übertragung eines komplexen Word-Dokuments.

Die *Synchronized Multimedia Integration Language SMIL* ist ein Ansatz zur Integration von Multimediadaten mit einem einfachen quelloffenen Autorensystem des W3C-Konsortiums. Viele Produktionen greifen aber auf bequemere, effektivere, aber leider auch komplexere Autorensysteme wie

Macromedia Director, *Totally Hip Technologies Livestage* oder *Macromedia Authorware* zu, die in proprietären Formaten gespeichert werden und deren langfristige Verfügbarkeit und Nutzbarkeit als problematisch eingestuft werden muß.

Zum wesentlichen Problem von Multimediaproduktionen wird es, dass die integrierten multimedialen Objekte für einen Umkodierungsprozeß am besten extrahiert und neu zusammengefügt werden. Die Möglichkeit einfacher Extraktion gehört aber nicht zu den üblichen Eigenschaften komplexer Multimediaprodukte, so dass ihre Umkodierung ein schwieriger Prozeß werden kann.

Viele Multimediaproduktionen hängen von bestimmten *Container*-Programmen ab, wie z.B. QuickTime, QuickTime VR, Real Audio, Real Video oder Windows Media Player (WMF). Zwar sind sowohl Quicktime wie Real Video für MacOS, MacOS X, Windows und für Linux zugänglich, ihre jeweils proprietären Definitionen sind trotzdem ein Warnzeichen unter dem Aspekt langfristiger Bestandssicherung.⁵

Für Multimediaproduktionen ist die Vorhersage für eine Bestandsicherung düster. Umkodieren ist vermutlich nur dann erfolgversprechend, wenn die Ausgangsdaten erhalten und verfügbar sind, wobei der Umkodierungsprozeß dem Aufwand einer Neukodierung nahe kommen kann. Ein anderer Weg mag über Emulation führen; dies ist freilich dadurch erschwert, dass Multimediaproduktionen häufig alle medialen Möglichkeiten eines Gerätes ausnutzen (die dann auch alle emuliert werden müssen).

⁵ Der MS Windows Media Player ist dagegen nicht einmal für Linux verfügbar.

Metadaten

Metadaten beschreiben Daten, Kontext, Inhalt und Struktur von digitalen Objekten, sowie deren Verarbeitungsgeschichte. Metadaten können ein zusammengefasstes Dokument (etwa durch Autor, Titel, Datum, Versions- oder Migrationsdaten) ebenso wie die einzelnen digitalen Bestandteile beschreiben. Zugriffsrechte oder Authentifizierungen gehören ebenso zu den Metadaten wie Hinweise auf den Zusammenhang mit anderen digitalen Objekten. Metadaten sind in diesem Sinne Zusatztexte, die programmiert interpretiert werden müssen. Ihre Zukunftssicherheit hängt von einer akzeptierten Standardisierung ab. Die technischen Anforderungen an ihre standardisierte ASCII-Textform sind dagegen gering.

Metadaten können durch Autoren ebenso wie durch Bibliotheken oder andere Sammelstellen erhoben werden. Sie erschließen digitale Objekte und vereinfachen die Suche nach ihnen. Sie dienen auch als eine Art Inhaltskompression und Verschlagwortung, wodurch sie die Dokumentenerschließung mittels automatisierter Suchverfahren erleichtern. Institutionelle Angebote mittels Metadaten erhöhen die Sichtbarkeit digitaler Angebote ebenso wie deren Zitierbarkeit. Perspektivisch dienen Metadaten der Erkundung des „deep net“ oder „deep web.“ Metadaten unterstützen und vereinfachen das Suchen in spezialisierten Datenbanken und mit spezialisierten Suchmaschinen.

• Dublin Core - Semantische Minimaldaten

Dublin Core ist ein Metadatenformat zur Beschreibung von Dokumenten und anderen Objekten im Internet. In seiner einfachen Version als *Dublin Core Metadata Element Set* besteht es aus 15 Datenfeldern. Alle Felder sind optional und können auch mehrfach vorkommen. Bei Bedarf werden sie

durch *Qualifiers* genauer spezifiziert. *Dublin Core* ist eine XML-Anwendung. Im Gegensatz zu MAB und MARC ist *Dublin Core* kein bibliothekarisches Metadatenformat, da es keine genauen Regeln für die Belegung der einzelnen Felder gibt, sondern eine Art „kleinsten gemeinsamen Nenner“ für den Austausch von Metadaten bietet.

1. *Title*: Titel
2. *Creator*: Urheber
3. *Subject*: Thema
4. *Description*: Kurzbeschreibung
5. *Publisher*: Rechteinhaber
6. *Contributor*: Mitautoren
7. *Date*: Datum
8. *Type*: Dokumententyp (erlaubt sind Collection, Dataset, Event, Image, Interactive Resource, Service, Software, Sound oder Text)
9. *Format*: Datenformat bzw. den Medientyp, auch MIME-Type
10. *Identifier*: Bezugsadresse für die Datei, z.B. URI/URL
11. *Source*: Quelle, von der die aktuelle Datei abgeleitet ist
12. *Language*: Sprache des Dokuments
13. *Relation*: Bezug zu einem Editionsprojekt
14. *Coverage*: zeitlicher oder geografischer Bezugspunkt für den Inhalt der Datei
15. *Rights*: Urheberrechtsangaben oder Verweis auf diese

Semantic Web

Mit dem *Semantic Web* soll das WWW um maschinenlesbare Daten erweitert werden, welche die Semantik der Inhalte formal festlegen. Das Konzept des *Semantic Web* geht auf einen Vorschlag von Tim Berners-Lee zurück.

Informationen sollen zusätzlich zu der für Menschen lesbaren Form auch formal, in einer für Maschinen bearbeitbaren Form repräsentiert werden,

damit Programme diese verarbeiten können. Das *Semantic Web* soll das WWW mit einem verteilten semantischen Netz überziehen. Die dazu nötige Annotation der HTML/XML-Seiten im Web kann z. B. mittels Repräsentationssprachen formaler Ontologien wie RDF oder dem darauf aufbauenden OWL erfolgen. Ob und wie freilich solche „formalen Ontologien“ erfolgreich und allgemein akzeptiert fixiert werden können, muß sich erst zeigen.

Nach Berners-Lee ist das *Semantic Web* eine Erweiterung des herkömmlichen Webs in der Informationen mit eindeutigen Bedeutungen versehen werden, um die Arbeit zwischen Mensch und Maschine zu erleichtern: „The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.“ (Tim Berners-Lee, James Hendler and Ora Lassila, *The Semantic Web*, 2001)

V. Emulation zur langfristigen Bereitstellung von Multimedialdokumenten und Programmen

Wie können alte Geräte und Software besser bereitgestellt werden? Durch Geräteerhalt oder durch Emulatoren? Eine naheliegende Idee mag es sein, Geräte und Software einer bestimmten technischen Entwicklungsstufe „einzufrieren“ und für spätere Nutzung bereit zu halten. Diese Idee kann als gescheitert angesehen werden. Sie funktioniert weder bei großen Organisationen noch bei Massenprodukten. Wer kann schon noch ein „Wordstar“-Dokument von einer 5,25“-Floppy lesen? Das Lesen einer älteren Floppy ist nur mit einem entsprechenden Gerät (Floppy, Controller, Schnittstelle, Treiber) möglich. Ist der Datenstrom dieser Floppy einmal in den Hauptspeicher übertragen worden, kann er auf einem modernen Rechner weiterverarbeitet werden – wenn dessen Software den Inhalt der Floppy interpretieren kann.

Für diesen Softwareteil der Bestandserhaltung bieten sich Softwareemulatoren an. Emulatoren sind seit der Urzeit der Großrechner bekannt, da große Organisationen nur dann einen Wechsel zu einer neuen Hardware vornehmen wollten (oder konnten), wenn sie die Daten und Programme der vorherigen DV-Anlage zumindest zeitweise weiter nutzen konnten.

Auch im PC-Bereich sind Emulatoren bekannt. Die Firma Apple hat den Übergang von der ursprünglichen Prozessorbasis des Macintosh, dem Motorola $\mu 68000$, auf ihre derzeitige Prozessorbasis, den von IBM sowie Motorola gebauten PowerPC mit bewundernswertem Erfolg bewältigt. Einen weiteren erfolgreichen Schritt hat Apple mit der Integration eines Emulator für das ursprüngliche MacOS Betriebssystem (in Form von OS 9.2) in das völlig anders aufgebaute, auf BSD-Unix beruhende, Betriebssystem OS X bewältigt. Dennoch hat es bei diesen Übergängen auch Verluste

gegeben – nicht viele, aber eben doch beim einigen Programmen, die sich nicht an die relativ strikten Vorgaben der *Apple Programming Guidelines* gehalten haben.

Microsoft hat kürzlich das Programm *VirtualPC* erworben, das es erlaubt, eine Intel PC-Architektur in einigermaßen hinreichender Arbeitsgeschwindigkeit auf einer Apple PowerPC-Architektur zu emulieren. Damit ist die Übertragung eines Microsoftbetriebssystems neben der Emulation vieler Hardwareschnittstellen auf eine „wesensfremde“ Macintosh-Architektur möglich. Tatsächlich lassen sich so unterschiedliche MS-Betriebssystemversionen wie DOS 2.1, Windows 3.11, Windows 2000, Windows NT oder Windows XP auf einem Mac emulieren. VirtualPC läßt sogar die Emulation eines Linuxbetriebssystems für die Intelarchitektur auf dem Mac zu.

Microsoft zeigt offensichtlich aber auch Interesse daran, ältere Windows-Versionen auf Intel-Architekturen unter neueren Windows-Versionen emulieren zu können, was für einzelne CD-Produktionen tatsächlich die einzige Möglichkeit bedeutet, sie weiterhin nutzen zu können.

Es gibt andere Emulatoren für ältere Geräte und Betriebssysteme. Eine eigene Szene zur Emulation hat sich im Kontext der Computerspiele entwickelt, wobei die technischen Anforderungen weit über die Textverarbeitung hinausgehen und in Bezug auf Grafik und Audio eine solide Basis für die Emulation von Multimediaprogrammen und -daten bereit stellen. Die Spieleemulation wird im Wesentlichen von Hobbyisten häufig im *Open-Source*-Umfeld vorangetrieben.

Emulatoren stoßen an Grenzen, wenn es gilt proprietäre Betriebssysteme ohne Kenntnis des Quellcodes oder ohne rechtliche Nutzungsmöglichkeit

evt. vorhandener Programmtexte nachzubilden. In dieser Hinsicht haben Apple und Microsoft mit ihren Emulatoren eine glückliche Ausgangslage, wenngleich das VirtualPC-Produkt auch ohne Microsoft-Unterstützung bis zur Version 6 eine reife Leistung bot.

Fraglos bestünde eine wesentliche bessere Ausgangslage für eine Emulation, wenn auf quelloffene Betriebssysteme wie Linux zurück gegriffen werden kann.

Eine andere Form der Emulation konkreter Geräte und Betriebssysteme geschieht durch den Einzug einer virtuellen Abstraktionsebene, wo ein Rechner in Softwareform angeboten wird, der selber auf einer oder eher vielen unterschiedlichen Geräte- und Betriebssystemplattformen implementiert ist. Die *Java Virtual Machine* JVM ist ein bekanntes Beispiel dieser Vorgehensweise. Java-Programme werden von diesem Softwareemulator ausgeführt. Da die JVM von vornherein ein Emulator ist, kann sie vergleichsweise leicht für unterschiedlichste Geräte und Betriebssysteme implementiert werden und sie ist auch auf sehr vielen solchen aktuellen Plattformen vorhanden.

Die JVM böte eine sehr gute Bestandshaltungsgarantie, wenn sie von allen beteiligten Parteien (Gerätehersteller, Betriebssystemhersteller, Programm-entwickler und Content-Ersteller, sowie Kunden und Bibliotheken) hinreichend unterstützt würde. Als ein gewisses Hindernis erweist sich die zögerliche Haltung der Firma SUN, ihre Rechte an Java einem offenen Standardisierungsgremium zu übergeben. Positiv ist jedoch die Freistellung des SUN Betriebssystems Solaris unter eine Open-Source-Lizenz zu sehen. Dennoch könnte eine klare Javaorientierung jenseits der kleinlichen

Streitereien um Effizienz und Ästhetik der Sprachkonzepte allein aus dem Grund langfristiger Erhaltung digitaler Objekte sehr sinnvoll sein.

Andere, wenngleich ähnliche Wege wurden zur Bestandserhaltung digitaler Objekte vorgeschlagen. Der IBM-Forscher Raymond E. Lorie hat ein Konzept eines *Universal Virtual Computers* vorgestellt. Auch Jeff Rothenberg verfolgt ein Emulatorkonzept als Rekonstruktion zu erhaltender Bestände.

VI. Migrationspfade zur Bestandssicherung: Organisation, Zeithorizont, strategische Aspekte

Analoge Wandlungen zur Bestandserhaltung

Die Speicherung analoger Dokumente kennt viele Migrationsformen zur Bestandsicherung. Beispiele sind der Nachdruck, der Kupferstich von Bildvorlagen, die fotografische Wiedergabe oder die Xerokopie. Für diese Kopien sind bewährte Aufbewahrungstrategien bekannt. Es fragt sich, wie weit digitale Objekte nicht auch in diesen analogen Verwahrgang eingefügt werden sollten. Für die Speicherung statischer Daten bieten sich die Techniken des Ausdrucks, der Mikrofilmspeicherung oder die Langfristspeicherung auf besonders widerstandsfähigem Material an.

Analoges Speichern digitaler Dokumente

Computerausdruck

Der Ausdruck auf Papier ist eine tradierte Form, um digitalisierte Texte aufzubewahren. Die Vorteile eines Ausdrucks auf säurefreiem Papier müssen hier nicht weiter erläutert werden. Problematisch kann freilich der Umfang ausgedruckten Textmaterials werden. Die CD-Textesammlung „Gutenberg.de,“ die ja in komprimierter Form eine deutschsprachige Literatursammlung mit vielen tausend Bänden enthält oder die DVD-Version des Großen Brockhaus oder der Encyclopedia Britannica mag dies verdeutlichen.

Neben der Kopie mittels normaler Drucker können auch höhere Auflagen mittels Computer-to-Plate-Verfahren erzeugt werden. Dies führt zum

Publishing-on-Demand. Eine Speicherung multimedialer Daten ist so freilich nicht möglich.

COM (Computer Output on Microfilm)

Eine klassische Alternative zum Papierausdruck bieten Mikrofilme, die derzeit bei idealer Lagerung durchaus 100 Jahre halten sollen. Kinofilme erreichen dieses Alter im Allgemeinen nicht, analoge Magnetbänder ebenfalls nicht.

Eine dynamische Speicherung durch Umkopieren, die bei den digitalen Speichern als zwingend erachtet wurde, kommt wegen der großen Qualitätsverluste bei analogen Kopien kaum in Frage. Unternehmensseitig ist die Datensicherung für Textdokumente auf Mikrofilm eingeführt, wobei die interessante Form *Computer Output on Microfilm* (COM) ist. Mikrofilmplotter können in ein Rechnernetz eingebunden und so konfiguriert werden, dass Primärinformationen und Metadaten gemeinsam derart auf Film gesichert werden, dass eine Redigitalisierung über Mikrofilmscanner möglich wird.

Die weiter oben erläuterte Rosetta Stone-Technik ist eine weitere Form analoger Langzeitspeicherung.

Wenngleich die analoge Kopie derart durchaus eine dauerhafte HiTech-Variante der Speicherung darstellen kann, sind analoge Kopien für Multimediadokumente gänzlich ungeeignet, da ihre interaktiven Elemente zur Navigation oder zu anderen Zwecken mit analogem Material nicht (oder nur mit völlig unverhältnismäßigem Aufwand) realisiert bzw. simuliert werden können.

Digitale Wandlungen analoger Objekte zur Bestandserhaltung

Imaging ist die Umwandlung visueller Objekte in ein digitales Format – typischerweise TIFF. Beim derzeitigen Stand der Technik gibt es keine zuverlässigen Such- und Indizierungsverfahren für solche visuellen Dokumente.

Bei Text-Images besteht die Möglichkeit der programmierten Schrifterkennung. *Optical Character Reading* (OCR) ist die programmierte Erkennung geschriebenen Textes. Dieser ist mit Standardmethoden indizierbar und durchsuchbar. Die positiven Erkennungsraten auf Wortebene liegen bei guten Vorlagen und den besten Programmen oberhalb von 95% (der Rest sind *Rückweisungen* oder *Erkennungsfehler*). Dies erlaubt bei redundanten Texten eine gute Indizierung auf Dokumentenseitenebene (99% werden berichtet).

Monaurale und stereophonische Tonobjekte können in AIFF gewandelt werden, dem Tonformat der Audio-CD.

Digitalisierte Objekte können komprimiert gespeichert werden. Kompressionsverfahren werden naheliegenderweise meist nach ihrer Kompressionsleistung und nach ihrem Zeitverhalten beurteilt. Dies sind unter dem Aspekt langfristiger Speicherung oder Archivierung nachgeordnete Kriterien. Hier sind unter dem Aspekt langfristiger Nutzbarkeit quelloffene Verfahren wie z.B. (verlustfreies) GZIP oder BZIP2 gegenüber proprietären Verfahren zu bevorzugen. Für Audiodaten liegt das quelloffene FLAC vor.

Verlustbehaftete Verfahren sind im Einzelfall auf ihre Einsetzbarkeit unter dem langfristigen Aspekt mehrfacher Umkodierung zu betrachten. Auch hier sind quelloffene nicht-proprietäre Verfahren wegen ihrer langfristigen Perspektive zu bevorzugen (MP3, PNG oder Ogg Vorbis). Auch für JPEG

oder MPEG-1, MPEG-2 oder MPEG-4 und XviD liegen frei zugängliche Quellen für Referenzimplementierungen vor.

Digital-digitale Migrationen

Digitale Speichermedien sind technologisch und physikalisch gefährdet. Dem kann nur durch eine Strategie des Umkopierens auf jeweils aktuelle Medien begegnet werden. Die technologische Aktualität erhält den Zugriff auf den gespeicherten Bitstrom. Das gespeicherte Objekt bleibt damit physikalisch lesbar.

Das Ziel jeder Bestandssicherung muß es sein, stabile und langfristig verfügbare Speicherformate zu verwenden. Speicherformate veralten aber technologisch ebenso wie Speichermedien. Speicherformate müssen deshalb unter solchen Umständen umkodiert werden, um interpretierbar zu bleiben.

Neben den physikalischen Binärdaten und ihrer unmittelbaren Interpretation als Datenobjekte sind weitere Angaben zur Bestandssicherung, die Metadaten, zu erhalten. Diese Metadaten betreffen Angaben zur Katalogisierung wie Erstellungsort, -datum, zum Charakter des Objekts, zu seiner Verschlagwortung bis hin zu den neuerdings beliebten Angaben zu formalen Ontologien.

Trotz der prinzipiell möglichen Datenmigration besteht stets die Gefahr eines unbemerkten Verlusts, wie bei einer digitalen „stillen Post.“ Die Minimierung solcher Verluste verlangt bei jeder Migration eine sorgfältige Kontrolle und gegebenenfalls Überarbeitung. Aus dem vergleichsweise wenig aufwendigen Aufbewahren eines materiellen Artefakts wird somit die

permanente, aktive Pflege eines digitalen Bestands - mit derzeit unabsehbarem Aufwand und Risiken – und entsprechenden Kosten.

Realistisch gesehen wird der Migrationspfad hin zur digitalen Speicherung umfassend beschritten werden. Die Probleme der langfristigen Speicherung und Bereitstellung sind für Textdokumente und einfache Grafiken lösbar. Für multimediale Dokumente und alle Dokumente, die programmierte Aktivität verlangen, sind weitere Forschungs- und Entwicklungsarbeiten nötig – aber es gibt für multimediale Artefakte keine brauchbaren Alternativlösungen im nicht-digitalen Bereich.

Zusammenfassung und Empfehlungen

- Die Speicherung und Nutzung digitaler Objekte über vier bis acht Jahre scheint relativ problemlos und für entsprechend aufbereitete Text- und Bilddaten stehen Techniken zur zehn- und zwanzigjährigen Speicherung und Nutzung zur Verfügung. Dies setzt jedoch die Vorhaltung entsprechender Geräte und Software oder deren Emulation voraus.
- Zur Speicherung und Nutzung multimedialer Daten über hundert Jahre sind keine gesicherten Strategien erkennbar.
- Bei selbst bestimmten Daten bestehen Chancen auf eine sachgerechte, ökonomisch und organisatorisch vertretbare langfristige Speicherung und Nutzung. Bei fremd bestimmten Daten sind keine solchen umfassenden Wege gesichert. Die proaktive Sammlung digitaler Objekte ist, soweit möglich, zu fördern. Zu bevorzugen ist eine enge Kooperation mit den Autoren/Erzeugern multimedialer Objekte, um langfristig speicherbare Formate zu sichern. Bei multimedialen Dokumenten ist die zusätzliche Speicherung der Ausgangsmaterialien zu bevorzugen.
- Es gibt keine dauerhaften digitalen Speichermedien. Notwendigerweise muß Bestandserhaltung bei digitalen Objekten als dynamische Speicherung erfolgen. Programme und Daten müssen als Schutz vor Schäden an den Speichermedien umkopiert werden, Programme müssen gegebenenfalls auch umkodiert werden, um sie an neue Geräte und Betriebssysteme anzupassen.

- Umkopieren und vor allem Umkodieren zur Anpassung an neue Geräte und Betriebssysteme wird zu einem wesentlichen Kostenfaktor für die Digitale Bibliothek, wobei der Aufwand zur Umkodierung den des Umkopierens um ein Vielfaches übersteigen kann.
- Neben der Objektspeicherung ist die frühzeitige Erstellung und zeitnahe Pflege zugehöriger Metadaten von wesentlicher Bedeutung.
- Authentizität und Integrität von digitalen Objekten sind zu sichern. Erfahrungen, die z.B. mit dem „Dokumentenserver der Humboldt-Universität“ (edoc.hu-berlin.de) mit hunderten gespeicherter Dissertationen und Habilitationsschriften gewonnen wurden, sind entsprechend auszuwerten.
- Unter dem Aspekt der technologischen Erhaltung sind rechtlich fixierte und vollständig offen gelegte Standards proprietären Systemen und so genannten Industriestandards vorzuziehen. Quelloffene Software wirkt als vorausschauende Bestandserhaltungsstrategie. Bei Computerprogrammen ist der Zugriff auf den Quellcode höchst wünschenswert, um das Umkodieren zur Nutzung unter neuen Geräten und Betriebssystemen zu erleichtern. Open-Source-Varianten sind deshalb vorzuziehen und zu fördern.
- Digitale Objekte sind meist einfach zu kopieren. Deshalb ist für die Datensicherung eine geeignete Backup-Strategie zu entwickeln, die jedes Objekt mindestens dreifach, vorzugsweise auf getrennten Speichermedien speichert.

- Abhängig von der Bedeutung des zu bewahrenden Objekts sollte die Kopie auf unterschiedlichen Datenträgern, gegebenenfalls auch in analoger Form, erwogen werden (z.B. Ausdruck wichtiger Dokumente zusätzlich zur digitalen Speicherung)
- Der Einsatz physisch und technologisch langlebiger Speichermedien ist selbstverständlich geboten. Forschung und Entwicklung langlebiger Speichermedien ist zu unterstützen, da wegen des Zwangs zum regelmäßigen Umkopieren hier ein wesentlicher Kostenfaktor der digitalen Bibliothek liegt.
- Teurer als Umkopieren ist allerdings Umkodieren wegen technologischer Veralterung von Geräten und Betriebssystemen. Da es keine Garantie für den Bestand von Hardware und Betriebssystemen gibt, scheint die Entwicklung von Emulatoren für verwendete Hardware und Betriebssysteme zum technischen Baustein mit dem Ziel der Bewahrung und Nutzung digitaler Objekte der beste Ausweg. Dies wird zu einer wesentlichen Herausforderung für Forschung und Entwicklung. Zu fördern ist die Forschung und Entwicklung an Emulatoren für alle medialen Objekte, vor allem für interaktive Multimediaobjekte.
- Die reine Datenspeicherung und Datensicherung lässt sich von den originären Aufgaben der Sammlung, der Bestandserschließung und der inhaltlichen Bestandssicherung abtrennen. Insbesondere unter Nutzung breitbandiger Rechnernetze könnte sie auch an eine vertrauenswürdige außen stehende Stelle, wie etwa an ein Rechenzentrum übergeben werden. Ebenso sind unter dem Aspekt einer *Economy of Scale* Ver-

bundlösungen zur Datensicherung denkbar. Auch in diesem Kontext ist eine stärkere Integration von Bibliotheken und Rechenzentren zu begrüßen.

- Besondere Problematiken entstehen beim ausschließlichen Netzzugriff. Unter Langfristüberlegungen ist der Speicherung beim Verlag die Speicherung im Haus oder die Hinterlegung bei einer vertrauenswürdigen dritten Stelle unbedingt vorzuziehen.
- Bibliotheken und andere Sammelstellen geraten mehr und mehr in die Rolle des Informationsmaklers oder des Verlags. Der Aufbau von Dokumentenservern, der Wunsch vieler Wissenschaftler und deren Organisationen zum Eigenverlag (*self-archiving*), die Zurückhaltung bestehender Verlage im Zeitschriftengeschäft, aber auch technische Entwicklungen, wie die immer einfacher werden Formen des *Publishing-on-Demand*, werden diese Tendenzen verstärken. Die Bibliotheken und Sammelstellen sollten sich diesen Tendenzen aktiv stellen. Dies erfordert eine offene Diskussion aller Beteiligten und Interessierten.
- Eine langfristige Speicherung und Nutzung von kommerziellen Multimediaproduktionen, bei denen keine Programmquellen und Originaldaten vorliegen, scheint höchst fraglich. Es ist darauf hinzuwirken, dass die erzeugenden Stellen Programme, Programmtexte und Originaldaten wirksam aufbewahren und zum Bedarf des Umformatierens und Umkodierens zur Verfügung stellen.

- Neben Lizenzen zum Netzzugang ist sicher zu stellen, dass eine körperliche Speicherung im Haus auf eigenen Speichermedien, sowie deren Nutzung über die Dauer des Netzzugangs hinaus gesichert ist oder die erworbenen Daten auf vertrauenswürdigen Servern zugänglich gespeichert werden – bis sie nach Ablauf des Urheberschutzes nach Belieben sicher gespeichert werden.
- Rechtliche Regelungen können die technischen Möglichkeiten aushebeln. Dies gilt insbesondere bei Regelungen zum Schutz geistigen Eigentums, wie dem Umgehungsverbot eines „wirksamen Kopierschutzes“ (§95a UrhG), das angesichts beschränkter Haltbarkeit digitaler Speichermedien eine vorhersehbare „Verrotten“ der gespeicherten Daten gesetzlich vorschreibt. Es wäre zu klären, wie weit ein solches Verbot einer Kopie dem Produkthaftungsgedanken widerspricht, da sich eine langfristige Speicherung oder Archivierung nicht auf die gesicherte Kooperation einer herstellenden Firma verlassen kann.
- Eine schwierige Lage besteht bei Multimediaproduktionen, da eine Sicherungskopie für Software sehr wohl zulässig ist (§69d UrhG). Ebenso ist das Dekompilieren von Programmen unter bestimmten Einschränkungen zulässig. Wie weit Multimediaproduktionen in diesem Sinne Software sind, ist jedoch nicht endgültig geklärt (Multimediale Artefakte gruppieren sich typischerweise um Programme herum oder sind Teil eines Programms). Doch grundsätzlich sind solche Einschränkungen der Sicherheitskopie aus der Sicht langfristiger Speicherung und Archivierung politisch abzulehnen oder dahingehend zu präzisieren,

daß Maßnahmen zu langfristigen Nutzung und Speicherung nicht behindert werden sollen.

Referenzen

- Abrams, Stephen L., David Seaman, Towards a global digital format registry, World Library and Information Congress: 69th IFLA General Conference and Council, Berlin, 2003, http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf
- Bearman, D., Reality and chimeras in the preservation of electronic records. *D-Lib Magazine*, 5(4), April 1999.
- Benfell, P., An integrated approach to managing electronic records. *Records Management Journal*, 12(3), 2002, p. 94-97.
- Berners-Lee, Tim, James Hendler & Ora Lassila, The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, Mai 2001
- British Computer Society Computer Conservation Society, Aims of the Society, <http://www.bcs.org/sg/ccs/> und http://cedarsgw.leeds.ac.uk/ccs/ccs_info.htm (1.6.2004)
- Dörr, Marianne & Hartmut Weber: Digitalisierung als Mittel der Bestandserhaltung? Abschlußbericht einer Arbeitsgruppe der Deutschen Forschungsgemeinschaft, *Zeitschrift für Bibliothekswesen und Bibliographie*, Jg. 44, H. 1, 1997, S. 53ff.
- Harnad, S., Universal FTP archives for esoteric science and scholarship: a subversive proposal, in: Okerson, A. & O'Donnell, J. (Hrsg), *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing 1*, Washington DC: Association of Research Libraries 1995. Auch: <http://www.arl.org/scomm/subversive/toc.html>
- Granger, S. (October 2000) Emulation as a digital preservation strategy. *D-Lib Magazine*, 6(10).
- Greenstein, Daniel & Suzanne E. Thorin, *The Digital Library – A Biography*, Digital Library Federation Council on Library and Information Resources, Washington, D.C., 2003
- J.Hunter, S.Choudhury, "A Semi-Automated Digital Preservation System Based on Semantic Web Services", Joint Conference for Digital Libraries, JCDL 2004, Tucson, Arizona 7-11 June 2004;
- J. Hunter, S.Choudhury, "Implementing Preservation Strategies for Complex Multimedia Objects", The Seventh European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2003, Trondheim, August 2003

- ISO 15489-1:2001 Information and documentation-records management, part 1: general, International Standards Organization. Auch: http://www.arxivervalencians.org/documents/ISO_15489-1.pdf (1.6.2004)
- Kennedy, J. and Schauder, C., *Records management a guide to corporate recordkeeping*. Reading (Mass.) u.a.: Addison Wesley, 2nd ed. 1998
- Lorie, Raymond A., Long-Term Archiving of Digital Information, Yorkstown Heights u.a.: IBM, Research Report RJ 10185 (95059), 18. Mai 2001
- Lorie, Raymond A., A project on preservation of digital data. In: *RLG DigiNews* 5 Nr. 3, 2001; Auch: <http://www.ohio.rlg.org/preserv/diginews/diginews5-3.htm> (1.6.2004)
- Multiple Arcade Machine Emulator, MameWorld Mission Statement, <http://www.mameworld.net/mission.html> (1.6.2004)
- Management, appraisal and preservation of electronic records. Vol 1: Principles, Kew, Richmond, Surrey, England: Public Record Office, 2nd ed. 1999
- Management, appraisal and preservation of electronic records. Vol 2: Procedures. Kew, Richmond, Surrey, England: Public Record Office, 2nd ed. 1999
- McInnes, S., Electronic records: the new archival frontier? *Journal of the Society of Archivists*, 19 (2), 1998, p. 211-20.
- Odlyzko, A. M., The economics of electronic journals, in: Ekman, R. & Quandt, R. (Hrsg.), *Technology and Scholarly Communication*, Berkeley: Univ. of Calif. Press, 380-393, 1998. Auch: <http://www.press.umich.edu/jep/04-01/odlyzko.html>
- Ohst, Daniel, Einsatz elektronischer Signaturen und Zeitstempel für die Sicherung digitaler Dokumente, Diplomarbeit an der Humboldt-Universität zu Berlin, Institut für Informatik, 2003.
- Quenault, H. (August 2001) VERS implementation project at the Department of Infrastructure – Melbourne, Australia. *Records Management Journal*, 11(2), p. 71-82.
- Rothenberg, Jeff, Ensuring the longevity of Digital Documents, *Scientific American*, Vol. 272:1, pp42-47, Jan. 1995
- Rothenberg, J. (1998) Avoiding technological quicksand: finding a viable technical foundation for digital preservation. Council on Library and Information Resources URL:<http://www.clir.org/pubs/reports/rothenberg/contents.html>
- Shepherd, E. and Yeo, G., *Managing records: a handbook of principles and practice*. London: Facet Publishing, 2003

The State of Digital Preservation: An International Perspective. Conference Proceedings, Washington, D.C.: Council on Library and Information Resources, July 2002

Wallace, D. A., Metadata and the archival management of electronic records: a review. *Archivaria*, 1993, 36 p. 87-110.

Wallace, D. A., Managing the Present: Metadata as Archival Description. *Archivaria*, 39, 1995, p. 11-21.